UNIVERSITE D'AIX-MARSEILLE
Faculté d'Economie Appliquée
Laboratoire de Mathématiques Appliquées
et d'Informatique

THESE

pour obtenir le grade de

DOCTEUR de l'UNIVERSITE D'AIX-MARSEILLE

Spécialité

STATISTIQUE MATHEMATIQUE

Par

Papa NGOM

Sujet de la thèse :

CRITERES INFORMATIONNELS ET TESTS D'HYPOTHESES

Soutenue le 19 mai 1998, devant la commission d'Examen :

Mr	M. Boutahar	: M. de conf.	Examinateur
Mr	P. Cazes	: Professeur	Rapporteur
Mr	R. Davidson	: Professeur	Rapporteur
Mr	C. Deniau	: Professeur	Examinateur
Mr	P. Hammad	: Professeur	Directeur de thèse

Remerciements

Je tiens à remercier :

Monsieur le Professeur Pierre Hammad, Directeur du Laboratoire de Mathématiques Appliquées et d'informatique, pour avoir accepté de m'accueillir dans son laboratoire, d'avoir guidé mes premiers pas dans la recherche et de m'avoir conseillé et suivi avec constance dans mes travaux. La confiance qu'il m'a accordée et son aide désintéressée ont été inestimables. Je tiens à lui exprimer ici toute ma gratitude et ma reconnaissance.

Je remercie les Professeurs Pierre Cazes et Russell Davidson qui ont accepté, par leurs rapports de juger ce travail, et m'ont, par leurs remarques constructives, épaulé dans mes recherches. Qu'ils trouvent ici toute ma gratitude

Les professeurs Claude Deniau et Mohamed Boutahar qui ont bien voulu me faire l'honneur d'être membres du jury. Je leur suis très reconnaissant des conseils et suggestions pour l'amélioration du manuscrit et pour l'accueil aimable qu'ils m'ont toujours réservé.

Mes remerciements vont également à toute l'équipe du Laboratoire de Mathématiques Appliquées, pour leur gentillesse et leur soutient.

Un remerciement particulier à Ciré Lamine , Maktar, Malick, Khadim, Seydou, Sidy, Fatimata Gassama, Pierre, Awa Dieng, Mamadou Diop, Birame Samb pour toute l'amitié qu'ils ont toujours manifestée à mon égard.

J'adresse un remerciement spécial à ma famille à laquelle je dois toute ma vie. Qu'elle trouve ici toute mon affection et ma plus profonde reconnaissance.

A mes parents
A ma famille

Résumé

Cette thèse s'insère dans le cadre de la théorie de l'information et de la statistique, en proposant, au travers de mesures de divergence, d'effectuer des tests d'hypothèses dans le cadre des modèles paramétriques, mais également des tests d'adéquation et de sélection de modèle. En s'appuyant sur la notion de distribution généralisée d'ordre α, nous nous intéressons d'abord (lorsque $\alpha = 1$), à la comparaison, d'une statistique de test informationnel, fondé sur une mesure de divergence J, avec les tests classiques usuels à distance finie. Nous établissons que cette statistique de test admet une propriété de robustesse, dans le cas où l'échantillon considéré est issu d'une loi exponentielle. Nous proposons ensuite un test de choix, fondé sur une classe de mesures de divergence, qui s'appuie sur une règle de décision qui prend en compte le niveau de signification du test.

Une autre partie est ensuite consacrée à une inférence bayésienne de la distribution généralisée, lorsque l'ordre α tend vers plus l'infini, afin d'élaborer une technique de calcul qui se révéle particulièrement performant notamment lorsque le calcul direct d'estimateurs du maximum de vraisemblance est impossible.

Mots clés : Théorie de l'information, Estimation, Test d'hypothèses, Simulation par Monte Carlo, Analyse bayésienne.

Abstract

This thesis comes within the information theory and statistics, by proposing, through of divergence measures, to perform hypothesis tests in the context of parametric models, but also goodness of fit test and test of model selection. Leaning on the generalized distribution notion, we take first an interest in (when $\alpha = 1$), the comparison of founded upon the divergence measure J, with classical statistic tests, in finite sample. We obtain, that this statistic test is robust, in the case where the sample considered is descended from an exponential law. We propose after, a test of choice based on with a class of divergence measures with a rule of decision, which take test size into account.

Another part is consacreted to a bayesian inference, about generalized distribution, in the case where the order α goes to infinity, with intent to construct a practical alternative to compute maximum likelihood estimator in exponential families in cases where a direct derivation is impossible.

Key words : Information theory, Estimation, Hypothesis test, Monte Carlo simulation, bayesian analysis.

Table des matières

Chapitre 1

Introduction générale

Dans ce chapitre, nous présentons le contexte général de notre recherche. Nous commençons d'abord par préciser la problématique dans la section 1.1. Ensuite, nous présentons respectivement dans les sections 1.2 et 1.3, une remarque méthodologique dont nous nous sommes inspirée tout au long de nos travaux, et une déscription de la démarche que nous avons suivie. Nous finissons ce chapitre en pésentant dans la section 1.4, l'organisation de cette thèse.

1.1 Problématique

La théorie classique de l'information a été introduite par l'ingénieur mathématicien américain Claude Shannon, dans une publication mémorable en 1948 portant sur la théorie mathématique de la conservation, la transformation et la transmission par lignes de l'information. L'énoncé de son théorème fondamental avait ouvert la voie à de nombreux développements aussi bien théoriques que pratiques dans le domaine des techniques de communication mais également dans les processus de décision, de détermination, de reconnaissance de formes, de réseaux d'ordinateurs ou encore des systèmes experts. Rappelons toutefois que dès 1928,

un autre ingénieur en transmissions, l'américain R.V. Hartley avait suggéré de mesurer le degré d'incertitude d'une épreuve à k éventualités par la quantité $log\ k$. Cette présentation de l'incertitude montre une forte similitude avec la définiton de l'entropie de Shannon. Ces deux formulations sont d'ailleurs asymptotiquement identiques notamment dans le cas d'une répétition infinie de la même épreuve (c.f Yaglom).

Du point de vue de la statistique, la théorie de l'information s'intéresse plus particulièrement à la quantification et à l'estimation de l'information et à l'application de celle-ci à des problèmes d'analyse ou d'inférence statistiques. C'est précisément le cas - entre autres - de la contribution remarquable qui a été proposée par Jaynes en (1957), par le biais d'une procédure qui repose sur l'estimation du "*maximum entropique*". Vont également dans ce sens, les travaux menés par Mokkadem (1994a) qui propose le choix d'un estimateur fondé sur une mesure de proximité entre deux densités spectrales sur le cercle unité, afin d'élaborer une méthode pour tester une hypothèse simple d'un processus ARMA. J. Gebert et al. (1969) ont introduit une procédure constuite à partir de l'estimation d'un indice, obtenu en fonction de la distribution empirique, pour résoudre un problème de test d'adéquation. Dans cet ordre d'idées, S. Kullback en (1967), développe une méthode d'estimation de la quantité d'information, qui repose essentiellement sur le prinicipe du "*minimum d'information discriminante*" qui est une extension du célèbre principe "*d'ignorance ou d'indifférence*" de Laplace. Il a ensuite élargi l'étude de l'estimateur de cette mesure d'information, en proposant, dans le cadre de la théorie des tests, une application pour la résolution de certains types de problèmes liés aux tableaux de contingence, aux hypothèses linéaires, à l'analyse multivariée, ainsi qu'aux fonctions de discrimination linéaires.

Divers auteurs ont entrepris des études d'unification des mesures d'informa-

tion, notamment les mesures de divergence et d'entropie, avant de proposer une application dans le cadre de la théorie générale de l'estimation et des tests d'hypothèses.

Bien que pertinente, cette approche que propose en particulier Morales, Pardo, Menendez et Salicrù (1992, 1993, 1994, et 1995), qui repose sur une étude systématique de la grande majorité des différentes mesures d'information, soulève cependant des problèmes que nous pouvons résumer par les questions suivantes : dans quelle mesure peut-on comparer les tests issus des mesures d'information avec d'autres plus classiques que sont, par exemple, les tests de Wald, du Score ou du Rapport de Vraisemblance? En effet, très peu d'approches d'analyse (comme dans O. Vasieck (1976) et (P. Barbe (1990)) ont été menées en vue d'une comparaison entre tests informationnels et tests classiques dans le cadre des modèles paramétriques. Dans le cas précis d'un test de sélection, peut-on mettre en évidence, un critère obtenu à partir d'une mesure d'information, permettant d'élaborer une règle de décision en association avec le niveau de signification que l'on s'accorde?

L'objet primordial de cette étude est de tenter d'apporter des éléments de réponse à ces questions qui découlent des constats suivants :

- une relative absence de travaux sur la comparaison entre les tests issus des mesures d'information et les tests classiques ;
- insuffisance de critères pour les tests de sélection qui tiennent compte du degré de précision du choix du décideur, à privilégier tel modèle plutôt que tel autre.

La première question est traitée, dans le cadre des modèles paramétriques, à partir de la distribution généralisée d'ordre α qu'évoque Hammad (1987), dans le cas particulier où α prend la valeur 1. Nous nous plaçons dans un contexte où on a défini une famille de lois de probabilité, indicées par un paramètre de dimension

finie, contenant la vraie loi des observations. En se limitant à une mesure de divergence J, nous montrons en l'occurrence que la statistique de test associée, $\hat{J}_n$, est plus robuste que les statistiques de Wald, du Score et du Rapport de Vraisemblance ; en ce sens qu'elle est moins sensible aux variations du paramètre d'intérêt.

Pour mettre en évidence les résultats obtenus, dans le cadre des échantillons de petite taille, nous proposons, de manière analogue à Davidson et Mackinnon (1994), une procédure de simulation pour comparer les performances par rapport aux tests classiques usuels.

La seconde question s'intéresse à l'élaboration d'un critère de choix entre deux modèles. Parmi les critères les plus utilisés, on peut citer le cœfficient de détermination, l'AIC de Akaike (1993), et BIC de Schwarz (1978) qui est une modification du critère de Akaike, tout en présentant l'avantage d'être convergent. La démarche alternative que nous avons adoptée pour proposer un critère de sélection, s'inscrit dans le cadre de l'approche introduite par Vuong et Wang (1993), et présente contrairement à ceux plus habituels cités ci-dessus, la possibilité de préciser le niveau de signification associé au test de sélection retenu.

Par ailleurs, lorsque la valeur α de la distribution généralisée tend vers l'infini, nous proposons une démarche alternative à celle introduite par Hammad (1992) pour une interprétation de la distribution généralisée d'ordre α. Cette démarche repose sur la méthodologie de l'approche bayésienne, et permet dans ce contexte, de mettre en évidence une procédure numérique de résolution des équations du maximum de vraisemblance, dans le cas précis où la loi de l'échantillon considéré appartient à la famille exponentielle.

1.2 Remarque méthodologique

Dans notre cadre de travail, nous adoptons le principe ci-dessous, concernant l'estimateur du paramètre d'intérêt θ. Le choix de cet estimateur sera principalement guidé par une propriété de normalité asymptotique.

Ce principe suppose que si $\widehat{\theta}$ désigne l'estimateur du paramètre θ, il est alors convergent et vérifie :

$$\sqrt{n}\ (\widehat{\theta} - \theta) \xrightarrow{L} N[0,\ Q]$$

où Q est l'inverse d'une matrice inversible.

On remarquera au passage, que ce principe garde un caractère de généralité, dans la mesure où il est conforme à la plupart des estimateurs convergents habituellement utilisés, tels que ceux issus du principe du maximum de vraisemblance, ou encore ceux obtenus par maximisation d'une fonction objectif.

Toutefois, avant d'expliciter comment nous avons exploité le principe énuméré ci-dessus, nous allons d'abord introduire une motivation d'ordre méthodologique qui nous a guidé tout au long de ce travail.

Ainsi, lorsqu'on utilise un échantillon aléatoire de n observations, sous les conditions de régularité habituelles, on peut estimer l'information de Kullback-Leibler

$$K[f_\theta,\ f_{\theta_o}] = \int f(x,\ \theta) \log \frac{f(x,\ \theta)}{f(x,\ \theta_o)} dx$$

par la statistique $\widehat{K}[f_\theta,\ f_{\theta_o}]$ en posant:

$$\widehat{K}[f_\theta,\ f_{\theta_o}] = \int f(x,\ \widehat{\theta}) \log \frac{f(x,\ \widehat{\theta}}{f(x,\ \theta_o)} dx$$

On vérifie sans difficulté que cette statistique est distribuée asymptotiquement suivant une loi du khi-deux :

$$2n\widehat{K}[f_\theta,\ f_{\theta_o}] \xrightarrow{L} \chi^2_p$$

avec p degrés de liberté (p étant la dimension de l'espace des paramètres Θ), où $f(x, \theta)$ est la densité de la loi des observations, $\widehat{\theta}$ étant un estimateur convergent et asymptotiquement normal et θ_o représentant la vraie valeur du paramètre.

Par analogie avec l'approche suggérée ci-dessus, nous avons cherché à estimer une mesure de divergence J, avant d'examiner son comportement asymptotique, pour ensuite en donner une application dans le cadre d'une inférence statistique.

1.3 Démarche suivie

La démarche suivie dans nos travaux peut être résumée brièvement en six étapes :

1. notre hypothèse de départ est la théorie de l'information ou plus particulièrement les mesures d'information dépendant de deux lois de probabilité. Nous donnons ensuite la définition de la distribution généraliée ϕ_α pour à une loi donnée.
2. nous avons adopté un critère informationnel, noté J pour évaluer l'écart entre deux distributions $f(x, \theta)$ et $f(x, \theta_o)$ lorsque $\alpha = 1$. L'estimateur $\widehat{J}_n$ de la mesure J sera obtenu, en remplaçant θ par $\widehat{\theta}$, vérifiant la propriété de normalité asymptotique.
3. la distribution de $\widehat{J}_n$ est ensuite déterminée, puis nous avons cherché à définir une application dans le domaine de la théorie des tests d'hypothèses.
4. une comparaison de $\widehat{J}_n$ avec les statistiques de test classiques est proposée dans le cadre des petits échantillons, appuyée par une étude de simulation par Monte Carlo.
5. nous avons proposé une méthode de test d'adéquation fondé sur une autre mesure de divergence Δ_r, plus générale, qui considère J comme un cas particulier. Une extension est également proposée pour la résolution d'un problème de test de choix, entre deux distributions appartenant à un modèle paramétrique.
6. en dernier lieu, nous nous sommes intéressés ensuite à une autre interprétation possible de l'inférence statistique, à partir d'une approche bayésienne de la distribution ϕ_α.

1.4 Organisation de la thèse

Cette thèse est composée de huit chapitres, organisés de la manière suivante :

- ce présent chapitre décrit nos motivations et guides méthodologiques ainsi que l'organisation de la thèse ;
- la partie I, constituée de deux chapitres, est consacrée aux critères, distances et estimateurs fondés sur des mesures d'information. Dans le chapitre 2, nous donnons quelques rappels sur les principaux résultats concernant les mesures d'information généralisées. Le but ici est d'introduire certains concepts de la théorie de l'information à partir des distributions généralisées, telle que l'information et distribution d'ordre α. Dans le chapitre 3, nous présentons une statistique de test $\widehat{J}_n$ fondée sur une mesure de divergence. On exposera les principaux résultats obtenus, notamment le comportement asymptotique de cette statistique suivant l'hypothèse considérée. Nous proposons, dans le cadre de la théorie des tests, une application dans laquelle, cette statistique $\widehat{J}_n$ est utilisée pour tester une hypothèse nulle de forme explicite ($\theta = \theta_o$) et une autre de forme implicite ($r(\theta) = 0$).
- la partie II est constituée de trois chapitres. Elle est consacrée à une étude comparative de la statistique $\widehat{J}_n$ par rapport aux statistiques de Wald, du Score et du Rapport de vraisemblance. Nous essayons, en nous fondant sur une série d'expériences par le biais de simulations de Monte Carlo, de se faire une idée du degré de performance en terme de puissance du test basé sur $\widehat{J}$. On propose également une méthode de test d'ajustement et de test de choix entre deux modèles à partir d'une mesure d'information notée Δ_r.
- la partie III est composée de trois chapitres. Le chapitre 6 traite

d'une approche bayésienne fondée sur la distribution d'ordre α.
Dans le chapitre 7, nous décrivons nos conclusions et perspectives de recherche ; le chapitre 8 regroupant les parties annexes.

$\triangle$

Première partie

Distances et estimateurs fondés sur des quantités d'information

Chapitre 2

Mesures d'information et distributions généralisées

2.1 Introduction

Ce chapitre vise à donner un aperçu de la théorie de l'information, en précisant certains aspects des mesures d'information sur lesquels nous nous appuierons tout au long de notre travail.

La théorie statistique de la communication, plus communément appelée théorie de l'information, est l'aboutissement des travaux d'un grand nombre de chercheurs -comme N. Nyquist, R.W.L Hartley, D. Gabor, etc. - sur l'utilisation optimale des moyens de transmission de l'information (téléphone, télégraphe, télévision, etc.). Rappelons que Claude E. Shannon, ingénieur chez Bell Téléphone, fût le premier a proposé, dans le cadre de cette théorie, une étude systématique, à travers un exposé synthétique (1948). L'idée fondamentale consiste à considérer l'information comme devant être transmise par le biais d'un canal (ligne téléphonique, ondes hertziennes). On est donc ramené à étudier trois situations :

- (a) d'une part on s'intéresse à la quantification de l'information au travers d'une série de mesures d'information (entropie, divergence etc.)

- (b) d'autre part, les propriétés liées aux canaux (équivoque, transinformation, capacité, etc.)
- (c) enfin les relations qui existent entre l'information à transmettre et le canal employé, en vue de l'utilisation optimale de celui-ci.

Dans le cadre de notre travail, nous nous intéressons uniquement au premier point souligné ci-dessus, à savoir le concept d'évaluation de l'information proprement dite. Nous chercherons ensuite une extension de ces mesures d'information en liaison avec la théorie des tests. Notons au passage que les concepts de base de la théorie de l'information, en raison des notions générales qui les sous-tendent font l'objet de plusieurs tentatives d'application dans des disciplines aussi diverses que les mathématiques, l'informatique, l'économétrie[1], la biologie et les sciences sociales[2].

2.2 Information et quantité d'information

Une information désigne, par définition, un ou plusieurs éléments parmi un ensemble fini d'évènements possibles. Pour synthétiser, considérons une expérience aléatoire qui peut être résumée par une variable aléatoire X prenant des valeurs discrètes $x_1, \ldots, x_n$ avec des probabilités $p_1, \ldots, p_n$ telles que $(P(X = x_i) = p_i,$ pour tout $i)$. Une question fondamentale est de savoir quantifier, par une évaluation plus ou moins précise, la notion d'information associée à cette expérience. Dans ce contexte, sur un plan purement pratique, une information sera d'autant plus intéressante qu'elle diminue davantage le nombre de possibilités ultérieures.

Une caractérisation de la mesure de l'information suite à l'analyse de cette expérience repose sur le choix approprié d'une fonction de base. Ainsi, sous l'angle de vue de la mécanique statistique, la quantité d'information apportée par cette expérience peut être mesurée par une expression fondée sur une fonction des pro-

1. C'est Davis (1941) qui a introduit le premier la théorie de l'information dans la littérature économétrique. Voir également la contribution en économétrie des travaux de G. Tintner, H. Theil, T. Sawa, E. Maasoumi, H. White, R. Klein, P.M Robinson, J. Segupta et A. Zellner.

2. Voir Massoumi (1988b)

babilités, que nous noterons par $g(p)$ avec $p = (p_1, \ldots, p_n)$ et telle que :

$$si \quad \exists\, k \;/p_k = 1, \quad \text{alors on a nécessairement} \quad g(p) = 0$$

$$\text{car} \quad p_i = 0 \quad \forall i \neq k$$

Ce qui signifie que la réalisation de x_k n'apporte aucune information.

Il est donc tout à fait naturel de considérer $g(p)$ comme une fonction décroissante de p. On suppose donc qu'elle doit satisfaire à la condition ci-dessous :

$$\begin{cases} g(1) = 0 \\ et \\ g(0) \longrightarrow +\infty \end{cases} \tag{2.1}$$

C'est dans cet ordre d'idées, que des auteurs comme Hartley (1928), Erdos (1946) ou encore Rényi (1961) - parmi tant d'autres - ont suggéré de définir g par une expression relativement simple par exemple en posant :

$$g(p) = \log(\frac{1}{p})$$

Rappelons que cette fonction doit également satisfaire à deux principes de base :

- principe d'additivité,
- principe de décroissance monotone,

qui sont résumés par le lemme fondamental suivant :

Lemme.

Soit $g(n)$ une fonction additive définie pour $n = 1,2,\ldots$.

Si elle vérifie en outre :

1. $g(nm) = g(n) + g(m)$; (additivité)
2. $\lim_{n\to\infty}[g(n+1) - g(n)] = 0$; (décroissance monotone)

alors $g(n) = a \ln n$ où a désigne la constante liée au choix de la base du logarithme.

2.3 Mesures d'information classiques

2.3.1 Systèmes d'axiomes et notion d'entropie

La quantité d'information, notée H, qu'on espère obtenir, à partir d'une expérience conduisant à n résultats, peut être définie par l'espérance mathématique de la fonction g définie précédemment.

On pose alors :

$$\begin{cases} H(p) = E(g(p)) = \sum_{i=1}^{n} p_i g(p_i) \geq 0 \\ p = (p_1, \ldots, p_n) \end{cases} \tag{2.2}$$

Dans le cas précis où la fonction g est définie par $g(p_i) = -\log p_i$, l'information $H(p)$[3] est connue sous le nom de mesure d'entropie de Shannon (1948).

Ainsi l'entropie[4] est en quelque sorte une mesure d'incertitude, de désordre ou d'imprécision, associée à une distribution ou à une variable aléatoire (ie une expérience). Historiquement, l'entropie relève d'un concept hérité de la mécanique statistique. C'est après qu'on a pu établir un lien avec le concept physique de l'entropie qui, elle, a été définie dans le cadre de la thermodynamique par le " *second principe de la termodynamique* ", qui trouve son origine dans le célèbre mémoire de Sadi Carnot intitulé " *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance*[5].

Nous considérons, en ce qui nous concerne, le concept d'entropie , définie sous l'angle statistique, c'est-à-dire, un formidable outil de mesure permettant de décrire un phénomène en utilisant le calcul des probabilités pour traduire le caractère aléatoire qui lui est attaché.

Rappelons toutefois que le rôle fondamental de l'information de Shannon, de Rényi et tant d'autres, ne doit en aucun cas masquer celui non moins important de

3. Notons que moyennant quelques conditions de convergences et de passage à la limite, on peut étendre cette mesure au cas multidimensionnel ou au cas continu en utilisant des intégrales et des densités.

4. Précisons cependant qu'il existe d'autres approches de l'entropie qui reposent notamment sur une série de systèmes d'axiomes (Fadeev 1957) voir également White (1992) pour une discussion détaillée sur la définition des mesures de probabilité dans les espaces continus.

5. Paris en 1824 (cf. Carnot)

l'information de Fisher, premier exemple de mesure d'information qui ait été étudié en Statistique Mathématique et plus précisément en théorie de l'estimation. En dépit de l'indépendance a priori de ces deux théories, on peut montrer, dans certains cas, les liens étroits qui existent entre le concept de Fisher et celui des autres mesures d'information[6]. La quantification de l'information qu'il propose représente sans nulle doute l'une des mesures d'information les plus fréquemment utilisées dans l'analyse statistique. Elle est en effet d'une remarquable utilité, précisément dans le traitement de problèmes spécifiques liés à la théorie de l'estimation ou à la théorie générale des tests.

2.3.2 Distance et divergence entre distributions

Dans de nombreux problèmes, l'inférence statistique se résume en une recherche de mesure, ou d'un critère permettant de porter un jugement sur le degré de proximité de deux (ou plusieurs) distributions de probabilité.

Ainsi, la notion de " *distance probabiliste* " entre deux distributions qui a été introduite initialement par Jeffreys, reste très souvent utilisée en théorie de l'information, sous le nom de *mesure de divergence.* Ainsi, si p et q sont deux distributions de probabilité, cette distance est définie par la fonction symétrique et non-négative suivante :

$$J[p, q] = \sum_i (p_i - q_i) \log \frac{p_i}{q_i}$$

D'autre part, lorsque la distribution p est donnée, il est tout à fait naturel de pouvoir mesurer, de manière quantitative, l'impact résultant de la substitution de p par une autre distribution q. C'est dans cet optique que Kullback et Leibler (1951) ont défini le gain d'information obtenu par remplacement de p par q, en

6. Pardo et al. ont montré la relation fondamentale qui apparait entre la mesure d'information de Fisher et toute une classe de mesures généralisées connues sous le nom de R-divergences dans une publication intitulée " *Generalized Jensen difference divergence measures and Fisher measure of information* " Kybernetes, Vol. 24 N 2, 1995 pp.15-28

considérant la quantité[7] :

$$K[p, q] = \sum_i p_i \log \frac{p_i}{q_i}$$

Cette mesure de divergence proposée par Kullback-Leibler - encore appelée divergence directe - constitue, en raison des interprétations intéressantes qu'elle engendre, une mesure privilégiée en statistique inférentielle. Cependant, dès lors qu'il s'agit de considérer une mesure de proximité entre distributions, il est important de noter qu'il existe par ailleurs une quantité de mesures de divergence qui permettent d'obtenir des résultats performants suivant le type de problème étudié, comme nous le verrons dans la section suivante.

7. La quantité d'information discriminante au sens de Kullback peut être regardée comme une mesure naturelle dans quelques cas particuliers importants comme l'exemple de deux lois normales $n-$dimensionnelles ayant des matrices de variance-covariance scalaires :
$P_1 = N(m_1, \sigma_1^2 Id)$, $P_2 = N(m_2, \sigma_2^2 Id)$
On a alors:

$$K[P_1, P_2] = n \log \frac{\sigma_1}{\sigma_2} - \frac{n}{2} + \frac{n\sigma_2^2}{\sigma_1^2} + \frac{||m_1 - m_2||^2}{2\sigma_1^2}$$

Ce qui laisse apparaître une distance entre moyennes et mesure de proximité entre variances

2.4 Mesures d'information généralisées

Il existe évidemment plusieurs fonctions, non-négatives, susceptibles de vérifier la condition (2.1). A partir de là, plusieurs définitions ont été proposées conduisant ainsi à toute une panoplie de mesures d'information. Il apparaît tout à fait utile, peut être même nécessaire de mener des études de synthèse, dont le but principal est de proposer des mesures de généralisation, donc d'unification des mesures d'information classiques. C'est ainsi qu'en se fondant sur une démarche méthodologique initialement introduite par Schützenberger[8], Rényi a fait un exposé synthéthique permettant d'étendre les mesures de Hartley, Shannon, etc.

C'est dans cet esprit qu'il propose une entropie d'ordre α définie par la relation suivante :

$$H_\alpha(p) = \{\alpha - 1\}^{-1} \log \sum_{i=1}^{n} p_i^\alpha$$

Un peu plus tard, une autre approche de la mesure d'entropie de degré r a été étudiée par Havrda et Charvat (1967) et définie comme suit :

$$H_r(p) = \begin{cases} \{r-1\}^{-1}[1 - \sum_{i=1}^{n} p_i^r] & si \quad r \neq 1 \\ -\sum_{i=1}^{n} \log p_i & sinon \end{cases}$$

Précisons que beaucoup d'autres mesures d'entropie ont été développées par la suite dans la littérature statistique.

Et c'est en se basant sur ces classes d'entropie, qu'une multitude de mesures de divergence, pour une variable aléatoire continue - ou pour une variable discrète - ont été développées plus précisément dans des domaines où interviennent le concept de variabilité dans un système où entre individus d'une même population. On remarquera les contributions - entre autres - de Jeffreys (1946) qui définit les J-divergences, de Rényi (1961), qui donne la première généralisation de l'information discriminante au sens de Kullback-Leibler, de Csiszar (1967) qui introduit les ϕ−divergences, de Burbea et Rao (1982) qui proposent les R-divergences, de Ta-

8. Schützenberger M. P "Contributions aux applications statistiques de la théorie de l'information", Inst. Stat. Univ. Paris (A) 2575, 1953.

neja (1989) qui étudie une extension de la généralisation des J-divergences ainsi que les R-divergences.

Dans le même ordre d'idées, soucieux de présenter une étude d'unifiaction des mesures de divergences, Morales, Pardo, Salicrù et Menendez (1993-1994) proposent une généralisation fondée sur deux fonctions, appelées "(h, ϕ)-*divergence* ", qui inclut les mesures citées précédemment comme cas particuliers.

Dans les limites de notre travail, et en raison du rôle important qu'elle peut jouer dans certaines applications, nous nous intéressons, principalement, dans le paragraphe qui suit, à un type de mesure d'information définie par le biais d'un paramètre scalaire et dans le contexte où les distributions considérées sont supposées incomplètes.

2.5 Information d'ordre α et disrtibution incomplète

2.5.1 Information d'ordre α

Une nouvelle approche de la caractérisation de l'information par le biais d'une idée heuristique de la notion de distribution incomplète a été introduite par A. Rényi et qui est en relation avec l'observabilité d'une épreuve. L'idée fondamentale consiste, lorsqu'on dispose d'un espace probabilisé $(\Omega,\mathcal{A},\mathcal{P})$, à considérer une variable aléatoire dite incomplète, qui se distingue d'une variable aléatoire ordinaire par le fait qu'elle n'est pas nécessairement définie pour tout événement ω de Ω. Dans ce sens, on peut dire que les variables aléatoires ordinaires constituent un cas particulier des variables aléatoires incomplètes.

Ainsi, si X est une variable aléatoire incomplète, prenant les valeurs $x_1,\ldots,x_n$ associées aux probabilités $p_1,\ldots,p_n$, on a:

$$\sum_{k=1}^{n} p_k \leq 1 \text{ et non nécessairement } \sum_{k=1}^{n} p_k = 1$$

Se faisant, à toute distribution incomplète $P = (p_1,\ldots,p_n)$, on peut faire correspondre une distribution complète, c'est à dire ordinaire $P' = (p'_1,\ldots,p'_n)$, en introduisant un facteur normatif.

On a alors:

$$p'_k = \frac{p_k}{\sum_{i=1}^{n} p_i}$$

Pratiquement, une variable aléatoire incomplète peut être interprétée comme une grandeur liée au résultat d'une épreuve - mais qui n'est pas définie pour tout résultat de l'épreuve - qu'avec une probabilité $\sum_{i=1}^{n} p_i < 1$. La variable aléatoire complète correspondante, quant à elle, peut être regardée comme une variable aléatoire conditionnelle de X, sous la condition qu'on puisse observer le résultat de l'épreuve.

C'est dans cet ordre d'idée que Rényi a défini, à partir d'une distribution incomplète $P = (p_1,\ldots,p_n)$ quelconque, une mesure d'information d'ordre α, par

la quantité

$$I_\alpha(P) = \frac{1}{1-\alpha} \log \left[\frac{1}{\sum_{k=1}^{n} p_k} \sum_{k=1}^{n} p_k^\alpha \right] \quad \alpha \neq 1$$

Le passage à la limite ($\alpha \longrightarrow 1$) redonnant l'information d'ordre 1, ou information de Shannon.

2.5.2 Distribution généralisée

En se fondant sur le concept de distribution incomplète évoquée au paragraphe précédent, P. Hammad propose une étude systématique d'une nouvelle classe de distributions généralisées[9](ou distributions d'ordre α).

De manière générale, à partir d'une densité f d'une loi complète donnée, on s'intéresse à la fonction f^α ($\alpha > 1$), qui lui est associée. On peut alors définir une densité complète ϕ_α, à partir de f^α, en introduisant un facteur de normalisation K défini par:

$$\begin{cases} \phi_\alpha = K f^\alpha \\ avec \\ K = \left(\int_R f^\alpha dx \right)^{-1} \end{cases} \tag{2.3}$$

On remarque évidemment que:

$$\lim_{\alpha \to 1} \phi_\alpha = f$$

La constante de normalisation K est une intégrale prise par rapport à la mesure de Lebesgue sur R, mais toute autre mesure ν, peut être considérée par exemple une mesure sur N dans le cas d'un modèle discret.
Cette nouvelle loi $\phi_\alpha dx$ sera appelée distribution généralisée d'ordre α, ou encore α-distribution.

L'interêt principal attaché à l'usage de la distribution généralisée, réside dans l'interprétation que l'on peut faire de l'ordre α. On peut envisager, comme le suggère P. Hammad (1987), une interprétation de α comme paramètre temporel, en étudiant un processus stochastique $X(t)$ à états continus et temps continus.

9. Voir Hammad (1987) p:137-158

Nous nous intéressons, en ce qui nous concerne, à une utilisation de la distribution ϕ_α dans la perspective d'une inférence statistique. Pour cela nous allons considérer deux cas :

(1) on supposera dans un premier temps que α prend la valeur 1, autrement dit que la distribution de densité complète ϕ_1 est identique à la densité traditionnelle f qui lui est associée. Il en sera ainsi dans les chapitres 3, 4 et 5,

(2) dans un deuxième temps, on tentera d'opérer un rapprochement entre α-distribution et analyse bayésienne, afin de proposer un algorithme pour la résolution numérique des équations de vraisemblance ; ce sera l'objet du chapitre 6.

$\triangle$

Chapitre 3

Test fondé sur la mesure de divergence J

3.1 Introduction

Les mesures de divergence occupent une place de choix en théorie de l'information mais également en analyse statistique notamment en théorie de l'estimation, des tests et région de confiance. La raison fondamentale vient en partie du fait que toutes ces mesures apparaissent comme un moyen quantitatif permettant d'évaluer l'idée de proximité entre distributions ou entre paramètres.

L'information discriminante au sens de Kullback peut être regardée comme un cas particulier d'une mesure plus générale appelée $\psi - divergence$, introduite par Csiszar (1967) et définie pour deux distributions arbitraires
$P = (p_i, \ldots, p_n)$ et $Q = (q_i, \ldots, q_n)$ par :

$$D_\psi[P, Q] = \sum_{i=1}^{n} q_i \psi(\frac{p_i}{q_i})$$

où $\phi : [0, +\infty[\longrightarrow \mathbf{R}$ est une fonction continue et convexe telle que :

$$0.\psi(\frac{0}{0}) = 0 \quad et \quad 0.\psi(\frac{x}{0}) = x \lim_{u \to +\infty} \frac{\psi(u)}{u}$$

Les $\psi - divergences$ qui reviennent le plus souvent dans la littérature, sont sans aucun doute, celles de Kullback-Leibler avec $\psi(x) = x \log x$, du $\chi^2 - divergence$ de Kagan avec $\psi(x) = 1 - x^2$, de Matusita avec

$\psi(x) = |1 - x^a|^{1/a}, 0 < a < 1$, de Balakrishman et Sanghvi avec $\psi(x) = \frac{(x-1)^2}{x+1}$, de Havrda et Charvat avec $\psi(x) = \frac{(x - x^s)}{1-s}$, $s \neq 1$, de Cressie et Read avec $\psi(x) = \frac{(x^{a+1} - x)}{a(a+1)}, \forall a \neq 0; a \neq -1$. D'autres exemples sont explicités dans Vajda (1989).

On remarquera cependant, qu'il existe d'importantes mesures d'information qui ne peuvent s'écrire sous la forme d'une $\psi - divergence$. C'est pourquoi, Menendez, Pardo, Morales et Salicrù proposent une extension de cette généralisation par une mesure appelée $(h, \psi) - divergence$, définie comme suit :

$$D_\psi^h[P, Q] = \sum_{a=1}^{A} \eta_a h_a [\sum_{i=1}^{n} q_i \psi_a(\frac{p_i}{q_i}) - \psi_a(1)]$$

où $h = (h_a)_{a=1,\ldots,A}$; $\psi = (\psi_a)_{a=1,\ldots,A}$ et pour $a = 1, \ldots, A$, ψ_a vérifie les conditions de la définition de Csiszar, les h_a sont non-décroissantes et continues sur $[0, \psi_a(0) - \psi_a(1) + \lim_{u \longrightarrow +\infty} \frac{\psi_a(u)}{u}]$ (cf. Theorem 9.1 in Vajda (1989)) et les η_a représentent des poids positifs. Ainsi, pour différentes fonctions de h_a et ψ_a, on obtient les mesures de Csiszar (1967), de Rényi (1961), de Sharma et Mittal (1977), de Taneja (1989), de Battacharyya (1946) etc.

Dans ce présent chapitre, notre but principal est de s'intéresser à l'estimation d'une mesure de $\psi - divergence$, avant d'étudier le comportement du test qui lui sera associé en rapport avec les statistiques de Wald, du Multiplicateur de Lagrange et du Rapport de Vraisemblance, lorsqu'on travaille dans le cadre des petits échantillons.

3.2 Modèle paramétrique

La problématique de la modélisation apparaît dès lors qu'on souhaite expliquer, décrire et analyser un phénomène réel. Un tel phénomème étant souvent complexe, il devient pratiquement impossible de l'appréhender dans sa totalité. Il est alors nécessaire de construire un résumé de la réalité permettant de l'étudier plus ou moins partiellement. C'est-à-dire que ce résumé ne prendra pas en compte toutes les caractérisques de la réalité, mais seulement celles qui semblent liées à l'objet de l' étude et qui ont une importance jugée suffisante.

Nous considérons, dans tout ce qui suit, le modèle statistique que nous définirons par :

(i) un modèle probabiliste :

$$\Phi = \{f(x, \theta), \theta \in \Theta\} \quad \text{et}$$

(ii) un modèle d'échantillonnnage :

$$X = (X_1, \ldots, X_n)^t.$$

Pour dresser une inférence sur le paramètre θ $(\theta \in \Theta)$, on considère une variable aléatoire X, admettant pour densité $f(x, \theta)$ pour tout $\theta \in \Theta$. On suppose que la famille de densités Φ vérifie les conditions de régularité définies ci-dessous

(i) *Soit* $A = \{x \in \Xi / f(x, \theta) > 0\}$ *ne dépendant pas de* θ *pour tout* $x \in A$, $\theta \in \Theta$, *les dérivées partielles suivantes existent et sont finies :*

$$\frac{\partial f(x, \theta)}{\partial \theta_i}, \qquad \frac{\partial^2 f(x, \theta)}{\partial \theta_i \partial \theta_j}, \qquad \frac{\partial^3 f(x, \theta)}{\partial \theta_i \partial \theta_j \partial \theta_k}, \quad i,j,k = 1, \ldots, p\,; \quad p = dim\ \Theta$$

(ii) *deux fonctions à valeurs réelles* $F(x)$ *et* $H(x)$ *telles que :*

$$\left|\frac{\partial f(x, \theta)}{\partial \theta_i}\right| < F(x), \qquad \left|\frac{\partial^2 f(x, \theta)}{\partial \theta_i \partial \theta_j}\right| < F(x), \qquad \left|\frac{\partial^3 f(x, \theta)}{\partial \theta_i \partial \theta_j \partial \theta_k}\right| < H(x),$$

où F est intégrable et $E[H(x)] < M < +\infty$, M indépendant de θ.

(iii) *La matrice d'information de Fisher*

$$I_X^f(\theta) = [E\{\frac{\partial \log f(x,\,\theta)}{\partial \theta_i} \frac{\partial \log f(x,\,\theta)}{\partial \theta_j}\}]_{i,j=1,\ldots,p}$$

est finie, définie positive.

Dans la section suivante, nous allons préciser la mesure de divergence sur laquelle nous nous appuyerons ainsi que son estimateur, avant d'en donner une application en liaison avec la théorie générale des tests.

3.3 Critère de décision informationnel basé sur la statistique $\widehat{J}_n$

Un tel critère est désirable pour apprécier le degré de proximité entre deux distributions données. Il existe cependant, comme nous l'avons rappelé précédemment, plusieurs mesures qui peuvent, de manière plus ou moins satisfaisante remplir ce rôle.

Nous nous limiterons ici au critère de la divergence J, fondé sur la somme de deux mesures de Kullback, pour deux distributions de probabilité données $f(x,\,\theta_1)$ et $f(x,\,\theta_2)$.

Ce qui revient donc à choisir la fonction $\psi(x)$ définie dans le paragraphe (3.1), en posant : $\psi(x) = x \log x$.

On a alors :

$$J[f(x,\,\theta_1),\,f(x,\,\theta_2)] = D_\psi[f(x,\,\theta_1),\,f(x,\,\theta_2)] + D_\psi[f(x,\,\theta_2),\,f(x,\,\theta_1)] \quad (3.1)$$

Cette mesure de divergence J - symétrique et non négative - est une pseudo-distance en ce sens qu'elle vérifie toutes les conditions necessaires pour l'existence d'une distance à l'exception de l'inégalité triangulaire. Elle peut être regardée comme un

outil permettant d'apprécier l'écart entre $f(x, \theta_1)$ et $f(x, \theta_2)$ sous l'angle informationnel.

Nous nous plaçons dans le contexte, où nous avons à évaluer la proximité entre θ et θ_o, la vraie valeur du paramètre, au travers des densités de probabilité correspondantes, $f(x, \theta)$ et $f(x, \theta_o)$.

En prenant appui sur l'expression définie en (3.1), on propose dans le cas présent, le critère de proximité qui prend la forme simple suivante :

$$J[f(x, \theta), f(x, \theta_o)] = \int_{\Xi} (f(x, \theta_o) - f(x, \theta)) \log \frac{f(x, \theta_o)}{f(x, \theta)} dx \qquad (3.2)$$

Ainsi, le manque d'information découlant du remplacement de θ_o par une autre valeur du paramètre θ, peut être apprécié au travers de la divergence J.

Une fois la mesure $J = J[f(x, \theta), f(x, \theta_o)]$ correctement spécifiée, il nous reste maintenant à lui associer un estimateur. Pour cela, nous pouvons considérer la statistique $\widehat{J}_n = n\widehat{J}$, obtenue en remplaçant θ, par son estimateur $\widehat{\theta}$. Le choix de l'estimateur du paramètre θ est fondé principalement sur le principe 1, adopté dans la section (1.2), qui est caractérisé par une convergence et un comportement asymptotiquement normal.

Cette propriété fondamentale nous permet, par la suite, d'étudier et de donner la loi asymptotique de $\widehat{J}_n$.

Dans le but d'étudier une inférence statistique à partir de $\widehat{J}_n$, nous avons choisi la théorie des tests asymptotiques. Rappelons que ces méthodes de test asymptotiques seront utilisées ici, lorsqu'on a spécifié un modèle paramétrique, c'est-à-dire lorsqu'on a défini une famille de lois de probabilité, indicées par un paramètre de dimension finie et contenant la vraie loi inconnue des observations.

Nous nous proposons de faire appel à $\widehat{J}_n$, pour résoudre un problème de test, dans le cadre d'un modèle paramétrique. Plus précisément, cette statistique sera utilisée dans les situations suivantes :

situation 1: (test d'une hypothèse explicite)

Considérons une série d'observations $x_1,\ldots,x_n$. En utilisant ces données, on estime le paramètre θ par $\widehat{\theta}$ et on procède au test suivant:

$$\begin{cases} H_o : \ \theta = \theta_o \\ H_1 : \ \theta \neq \theta_o \end{cases} \tag{3.3}$$

situation 2: (test d'une hypothèse implicite)

A partir des observations dont on dispose, on estime le paramètre θ par $\widehat{\theta}$ et on décide de tester simultanément plusieurs contraintes réelles:

$$\begin{cases} H_o : \ r(\theta) = 0 \\ H_1 : \ r(\theta) \neq 0 \end{cases} \tag{3.4}$$

où les fonctions $r_1,\ldots,r_k$ sont à valeurs dans R, dérivables et telles que la matrice $\dfrac{\partial^t r}{\partial \theta}$ soit de rang k, $\forall\ \theta \in \Theta$.

Le comportement asymptotique de la statistique $\widehat{J}_n$ est donné par le théorème suivant:

Théorème: 1

Soit $\widehat{J}_n$ l'estimateur de la divergence nJ obtenu en remplaçant $\theta = (\theta_1,\ldots,\theta_p)$ par un estimateur, $\widehat{\theta} = (\widehat{\theta}_1,\ldots,\widehat{\theta}_p)$ vérifiant le principe de normalité asymptotique.

Sous l'hypothèse nulle ($\theta = \theta_o$), on a:

$$\widehat{J}_n = nJ[f(x,\ \widehat{\theta}),\ f(x,\ \theta_o)] \xrightarrow{L} \chi^2_p$$

où $p = dim\Theta$

Preuve:

Considérons la fonction suivante:

$$\eta(\theta) = J[f(x,\ \theta);\ f(x,\ \theta_o)]$$

En appliquant le développement de Taylor de la fonction $\eta(\theta)$ autour de θ, on obtient :

$$\eta(\widehat{\theta}) = \eta(\theta) + \sum_{i=1}^{p}(\widehat{\theta}_i - \theta_i)\frac{\partial\eta(\theta)}{\partial\theta_i} + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(\widehat{\theta}_i - \theta_i)\frac{\partial^2\eta(\theta)}{\partial\theta_i\partial\theta_j}(\widehat{\theta}_j - \theta_j) + R_n \quad (3.5)$$

sous l'hypothèse $\theta = \theta_o$, on a : $\dfrac{\partial\eta(\theta)}{\partial\theta_i} = 0$ et les dérivées secondes de $\eta(\theta)$ sont égales à :

$$\begin{aligned}\frac{\partial^2\eta(\theta)}{\partial\theta_i\partial\theta_j} &= 2\int_{\Xi}\frac{\partial f(\theta)}{\partial\theta_i}\frac{\partial f(\theta)}{\partial\theta_j}\frac{1}{f(\theta)}dx \\ &= 2I_{ij}(\theta)\end{aligned}$$

où $I_{ij}(\theta)$ est le général de la matrice d'information de Fisher $I(\theta)$. En remplaçant $\dfrac{\partial^2\eta(\theta)}{\partial\theta_i\partial\theta_j}$ dans l'expression (3.5), puis en multipliant par n, on a :

$$n\ \eta(\widehat{\theta}) = \sum_{i=1}^{p}\sum_{j=1}^{p} n(\widehat{\theta}_i - \theta_i)I_{ij}(\theta)(\widehat{\theta}_j - \theta_j) + R_n$$

D'où l'on obtient :

$$\widehat{J}_n = nJ[f(x, \widehat{\theta}), f(x, \theta_o)] \longrightarrow \chi^2_p$$

3.4 Application aux tests d'hypothèses

On s'intéresse maintenant, dans la section suivante, à une procédure de test basée sur la statistique $\widehat{J}_n$, lorsque la loi des observations est issue d'un modèle paramétrique.

3.4.1 Cas d'une hypothèse nulle sous forme explicite : $\theta = \theta_o$

En se référant aux résultats obtenus à partir des propriétés asymptotiques de la divergence $\widehat{J}_n$, on peut chercher à établir une application dans le cadre de la résolution de problèmes liés aux tests d'hypothèses.

Considérons un échantillon $X = (X_1, \ldots, X_n)$ d'une variable X admettant une densité $f(x, \theta)$. La problématique consiste à confronter, sur la base de l'information dont on dispose, deux hypothèses complémentaires relativement à une même population ou à un même processus aléatoire. Dans ce contexte, les hypothèses à tester peuvent être formuler comme suit :

$$\begin{cases} H_o : & \theta = \theta_o \\ H_1 : & \theta \neq \theta_o \end{cases} \tag{3.6}$$

Ce qui est équivalent au test d'hypothèse défini par :

$$\begin{cases} H_o : & J[f(x, \theta), f(x, \theta_o)] = 0 \\ H_1 : & J[f(x, \theta), f(x, \theta_o)] \neq 0 \end{cases} \tag{3.7}$$

Dans ce cadre précis, nous considérons la statistique $\widehat{J}_n$ définie précédemment en (3.2) et la fonction de test donnée par :

$$\Phi(x_1, \ldots, x_n) = \begin{cases} 1 \; si \; \widehat{J}_n > c_\alpha \\ 0 \; sinon \end{cases}$$

La constante c_α est déterminée par la relation :

$$P[\widehat{J}_n > c_\alpha / H_o] = \alpha$$

avec

$$c_\alpha = \chi^2_\alpha$$

où $\chi_1^2(\alpha)$ représente la valeur de la loi du Khi-deux pour laquelle, la probabilité d'être dépassée est égale à α (α désignant le niveau de signification du test).

En guise d'illustration de l'utilisation de $\widehat{J}_n$, nous nous proposons d'examiner deux exemples, correspondant respectivement au cas où les observations sont issues d'une population normale où d'une loi exponentielle.

3.4.2 Exemples

(a) Cas gaussien

Soit X la variable (de dimension un pour simplifier) attachée à un échantillon de taille n, issu d'une population normale,

$$X \sim N[\theta, \sigma]$$

Nous nous intéressons au problème de décision défini en (3.3). Ainsi, si $f(x, \theta)$ est la densié de X , le test de la moyenne de cette loi (la variance est supposée connue), sera fondé sur la statistique $\widehat{J}_n$, qui prend alors une forme relativement simple :

$$\widehat{J}_n = n\widehat{J}[f(x, \widehat{\theta}), f(x, \theta_o)] = \frac{n}{\sigma^2}(\widehat{\theta} - \theta_o)^2 \longrightarrow \chi^2$$

La constante défissant la région critique est alors égale à $c_\alpha = \chi_1^2(\alpha)$.

Convergence du test

Sous l'alternative, $\widehat{\theta}$ converge vers une valeur $\theta_1 \neq \theta_o$; on a alors :

$$\widehat{J}_n = \frac{n}{\sigma^2}(\widehat{\theta} - \theta_o)^2 \longrightarrow +\infty \quad avec \quad n$$

c'est-à-dire :

$$P_n = Prob[\widehat{J}_n > c_\alpha / H_1] \longrightarrow 1$$

On en déduit que le test est asymptotiquement convergent au sens de Frazer.

(b) Cas de la loi exponentielle

Considérons une v.a.r X et $(X_1,X_2,\ldots,X_n)$ un échantillon i.i.d extrait de la loi de X dont la densité est définie par :

$$\begin{cases} f(x,\,\theta) = exp[-\theta x + \log(\theta)], \quad pour \quad x \geq 0 \\ \\ avec \quad \theta > 0 \end{cases} \tag{3.8}$$

On suppose qu'on désire résoudre le même problème de test formulé dans (3.3). Pour cela, un calcul simple de la divergence J montre que :

$$J[f(\theta),\, f(\theta_o)] = \frac{1}{(\theta\theta_o)}(\theta - \theta_o)^2;$$

On a :

$$\widehat{J}_n = n\widehat{J}[f(x,\,\widehat{\theta}),\, f(x,\,\theta_o)] \longrightarrow \chi^2$$

Propriété du test

Nous obtenons alors l'expression de la puissance donnée par la formule suivante, pour un niveau de signification égale à α :

$$P_n = P[\widehat{J}_n > c_\alpha / H_1]$$

et comme sous l'alternative $\widehat{J}_n \longrightarrow +\infty$ avec n, cela se traduit par la convergence asymptotique vers 1 de la puissance P_n.

3.5 Test d'une hypothèse nulle sous forme implicite : $r(\theta) = 0$

Nous commençons par rappeler la définition des tests classiques de base, dans le contexte de la théorie asymptotique, ainsi que les principales propriétés qui leurs sont associées. La méthode de construction de ces tests que sont le test de Wald, le test du multiplicateur de Lagrange (ou test du score) et le test du rapport de vraisemblance, est spécialement basée sur une fonction de vraisemblance. Le champ d'application de ces tests est particulièrement vaste, mais leurs principales propriétés sont fondées sur une justification essentiellement asymptotique.

Remarquons que ces tests sont très pratiques, notamment lorsqu'on dispose - ce que nous supposons dans ce qui suit - d'un modèle paramétrique défini par une famille de lois de probabilité, indicées par un paramètre de dimension finie.

3.5.1 Généralités

Nous nous plaçons dans le cas le plus général où l'espace des paramètres Θ ou le sous-espace Θ_o définissant l'hypothèse nulle H_o sont multidimensionnels (la situation ou le paramètre θ est un scalaire représente donc un cas particulier). On considère une variable aléatoire X et on désigne par $f(x, \theta)$ la densité de sa loi de probabilité.

La vraisemblance d'un échantillon $(X_1, X_2, \ldots, X_n)$ de la loi de X est donnée par :

$$\ell(x/\theta) = f(x_1, \theta), \ldots, f(x_n, \theta)$$

Notons $\mathcal{L}(x, \theta)$ la log-vraissemblance du modèle et par $\widehat{\theta}$ l'estimateur du maximum de vraissemblance de θ ($\theta \in \Theta$ et $\Theta \subset R^p$).

On a :

$$\mathcal{L}(x, \theta) = \sum_{i=1}^{n} \log f(x_i, \theta)$$

On suppose que les conditions de régularité habituelles sont satisfaites, et que la matrice d'information est définie par :

$$I(\theta) = -E\left[\frac{\partial^2 \log f(x, \theta)}{\partial\theta\partial\theta^t}\right] \tag{3.9}$$

qui est une matrice régulière et continue pour tout point θ de Θ. L'espérance mathématique dans l'expression (3.9) est prise par rapport à la loi de densité $f(x, \theta)$ et que les estimateurs convergents de $I(\theta)$ sont :

$$-\frac{1}{n}\frac{\partial^2 \mathcal{L}(x, \widehat{\theta})}{\partial\theta\partial\theta^t} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(x_i, \theta)}{\partial\theta\partial\theta^t} \tag{3.10}$$

ou encore :

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(x_i, \widehat{\theta})}{\partial\theta}\frac{\partial \log f(x_i, \widehat{\theta})}{\partial\theta^t} \tag{3.11}$$

On suppose que l'hypothèse nulle définie par le sous-ensemble Θ_o est une fonction de θ, et s'écrit sous forme implicite :

$$r(\theta) = 0 \tag{3.12}$$

et que la matrice des dérivées partielles $\dfrac{\partial r^t(\theta)}{\partial\theta}$ est de rang k ; $\forall\theta \in \Theta$; $k < p$.

Avant de présenter la forme que prend l'estimateur $\widehat{J}_n$ sous la contrainte $r(\theta) = 0$, nous exposerons d'abord les caractéristiques essentielles des tests asymptotiques.

3.5.2 Aperçu sur les procédures de tests asymptotiques classiques

(a) Test du Rapport de vraissemblance

Considérons l'échantillon formé d'observations i.i.d (X_i), $i = 1, \ldots, n$; un vecteur de paramètre $\theta \in \Theta$ (on se restreint au cas particulier où $\Theta = \Re^p$) avec une hypothèse nulle définie par la condition $r(\theta) = 0$.

Le principe du quotient de vraisemblance revient à étudier la statistique formée par :

$$\begin{aligned} R_n &= 2[\mathcal{L}(x, \widehat{\theta}) - \mathcal{L}(x, \widetilde{\theta})] \\ &= 2\log\frac{\ell(x, \widehat{\theta})}{\ell(x, \widetilde{\theta})} \end{aligned} \tag{3.13}$$

où $\ell(x, \widehat{\theta})$ et $\ell(x, \widetilde{\theta})$ désignent les fonctions de vraisemblance maximisées obtenues, respectivement, sans contrainte et sous l'hypothèse nulle.

En d'autres termes, $\widetilde{\theta}$ représente l'estimateur contraint de θ, obtenu par la méthode du maximum de vraisemblance :

$$\begin{cases} Max & \ell(x, \theta) \\ r(\theta) = 0, \end{cases}$$

alors que $\widehat{\theta}$ est l'estimateur obtenu par le maximum de vraisemblance sans contrainte, en résolvant

$$\begin{cases} Max & \ell(x, \theta) \\ \theta & \in \Theta \end{cases}$$

Intuitivement, si l'hypothèse nulle est fausse, alors la valeur de R_n sera " grande " (autrement dit la vraisemblance contrainte sera inférieure à la vraisemblance non contrainte), ce qui suggère le choix d'une région critique par rapport au critère $R_n > c$, où c est une constante déterminée par un seuil de signification α (il s'agit de la probabilité de rejeter H_o alors qu'elle est vraie). La région critique s'obtient en posant :

$$W_{R_n} = \{(x_1, \ldots, x_n)/Prob[R_n > c/H_o] = \alpha\}$$

Or, la valeur de la constante c ne peut être déterminée de façon précise que si nous connaissons la distribution exacte (petit échantillon) de la statistique R_n. Ce qui est rare en pratique, notamment dans les modèles très complexes. C'est pourquoi on fait appel à la théorie asymptotique qui fournit les bases de la construction d'un tel test, dès lors que les conditions normalité asymptotique et d'éfficacité de l'estimateur du maximum de vraisemblance sont remplies.[1]

Ainsi, sous certaines conditions de régularité, l'utilisation judicieuse de l'approximation par les séries de Taylor, montre que R_n est distribuée asymptotiquement suivant une loi du χ^2 avec k degré de liberté.

1. Voir Robert J. SERFLING: " *Approximation theorems of mathematical statistics* " *Wiley (1980).*

(b) Test du Wald

Le test du rapport de vraisemblance nécessite l'estimation du modèle à la fois sous l'hypothèse nulle et sous l'alternative. Le test de Wald, par contre, ne nécessite qu'une estimation du modèle non contraint et il est d'autant plus attrayant que le modèle contraint est, en général, difficile à estimer dans certains cas. Pour une meilleure compréhension intuitive de ce test, on va considérer brièvement le cas simple où l'hypothèse nulle est donnée par $\theta = \theta_o \in \Re$. Ce test est basé sur le fait que de gros écarts entre $\widehat{\theta}$ et θ_o signifient que les données rejettent l'hypothèse nulle. Se pose alors la question de savoir comment apprécier cette notion de distance selon un critère approprié?

On remarque en effet que deux échantillons distincts peuvent conduire à la même valeur $(\widehat{\theta} - \theta_o)^2$. Par conséquent cette distance doit donc être pondérée par ce qu'on appelle l'indicateur de courbure $C(\widehat{\theta})$, définie par:

$$C(\widehat{\theta}) = \Big(\frac{\partial^2 \log \ell(x,\, \widehat{\theta})}{\partial \theta^2}\Big)$$

La statistique de Wald est alors donnée par:

$$W_n = (\widehat{\theta} - \theta_o)^2 C(\widehat{\theta})$$

On peut généraliser ce résultat au cas $\theta \in \Re^p$; avec une hypothèse nulle symbolisée par la relation

$$r(\theta) = 0 \qquad (r : \Re^p \longrightarrow \Re^k)$$

Si G désigne, la matrice Jacobienne $(k \times p)$

$$G(\widehat{\theta}) = \Big(\frac{\partial r(\widehat{\theta})}{\partial \theta}\Big)_{1 \le i \le k,\quad 1 \le j \le p}$$

sous l'hypothèse nulle $(H_o : r(\theta) = 0)$, on a:

$$\sqrt{n}[r(\widehat{\theta}) - r(\theta_o)] \xrightarrow{L} N[0,\, G(\theta_o) I^{-1}(\theta_o) G^t(\theta_o)]$$

D'où on pose:

$$W_n = n r(\widehat{\theta})^t [G(\widehat{\theta}) I^{-1}(\widehat{\theta}) G^t(\widehat{\theta})]^{-1} r(\widehat{\theta})$$

On montre à partir de cette forme quadratique, que la statistique de Wald est distribuée asymtotiquement suivant une loi du khi-deux à k degré de libertés.

(c) Test du multiplicateur de Lagrange

Ce test provenant également d'une estimation par maximum de vraisemblance sous contrainte - conduite selon la méthode du Lagrangien - est souvent considéré comme équivalent au test du Score, où le score désigne les dérivées partielles premières

$$\frac{\partial \log \ell(x, \widehat{\theta})}{\partial \theta}$$

Contrairement au test de Wald, le test du multiplicateur de Lagrange nécessite l'estimation du modèle contraint. L'idée est la suivante: si l'hypothèse nulle est vraie, (ie les restrictions sont valides) alors l'estimateur du maximum de vraisemblance contraint $\widehat{\theta}_o$ sera " proche " de l'estimateur non contraint $\widehat{\theta}$. Et comme les estimateurs non contraints maximisent le *log* de la fonction de vraisemblance, ils satisfont l'équation :

$$S(\theta) = 0 \qquad \text{où} \quad S(\widehat{\theta}) = \frac{\partial \log \ell(x, \widehat{\theta})}{\partial \theta}$$

On peut donc formuler la statistique S_n, en posant :

$$S_n = \frac{1}{n} S(\widehat{\theta}_o)^t I^{-1}(\widehat{\theta}_o) S(\widehat{\theta}_o)$$

On démontre là également, que cette statistique suit, asymptotiquement une distribution

du khi-deux à k degré de libertés.

3.5.3 Test construit à partir de la divergence $\hat{J}_n$

De manière analogue au test du Rapport de vraisemblance, ce test est fondé sur la divergence entre les valeurs des estimateurs de la vraisemblance contrainte et non contrainte. Le test d'hypothèse nulle définie par (3.4) sera basée sur la statistique de la divergence

$$\hat{J}_n = nJ[f(\widehat{\theta}), f(\widetilde{\theta})]$$

La densité de probabilité de la loi asymptotique de $\widehat{J}_n$ est donnée par le théorème suivant :

Théorème : 2

Soit $\widehat{J}_n$ l'estimateur de la divergence nJ obtenu en remplaçant $\theta = (\theta_1, \ldots, \theta_p)$ par les estimateurs, $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)$ et $\widetilde{\theta} = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_p)$ vérifiant le principe de normalité asymptotique.

Sous l'hypothèse nulle H_o, on a :

$$\widehat{J}_n = nJ[f(\widehat{\theta}),\, f(\widetilde{\theta})] \xrightarrow{L} \chi^2_k$$

où k est le rang de la matrice des dérivées partielles de $r(\theta)$.

Preuve :

Soit θ_o la vraie valeur du paramètre.

Posons :

$$H(\theta) = \int f(\theta) \log \frac{f(\theta)}{f(\widetilde{\theta})} dx \quad et \quad G(\theta) = \int f(\theta) \log \frac{f(\theta)}{f(\widehat{\theta})} dx$$

Soit alors

$$\widehat{J} = J[f(\widehat{\theta}),\, f(\widetilde{\theta})] = H(\widehat{\theta}) + G(\widetilde{\theta})$$

Le developpement de Taylor appliqué aux fonctions $\log f(\widehat{\theta})$ et $\log f(\widetilde{\theta})$ au voisinage de θ_o donne :

$$\begin{aligned} \log f(\widehat{\theta}) &= \log f(\theta_o) + (\widehat{\theta} - \theta_o) \frac{\partial \log f(\theta_o)}{\partial \theta} \\ &\quad + \frac{1}{2} (\widehat{\theta} - \theta_o)^t \frac{\partial^2 \log f(\theta_o)}{\partial \theta \partial \theta^t} (\widehat{\theta} - \theta_o) + R_n^1 \end{aligned}$$

et

$$\begin{aligned} \log f(\widetilde{\theta}) &= \log f(\theta_o) + (\widetilde{\theta} - \theta_o) \frac{\partial \log f(\theta_o)}{\partial \theta} \\ &\quad + \frac{1}{2} (\widetilde{\theta} - \theta_o)^t \frac{\partial^2 \log f(\theta_o)}{\partial \theta \partial \theta^t} (\widetilde{\theta} - \theta_o) + R_n^2 \end{aligned}$$

Par différence on a :

$$\begin{aligned} \log \frac{f(\widehat{\theta})}{f(\widetilde{\theta})} &= (\widehat{\theta} - \widetilde{\theta}) \frac{\partial \log f(\theta_o)}{\partial \theta} \\ &\quad + \frac{1}{2} (\widehat{\theta} - \widetilde{\theta})^t \frac{\partial^2 \log f(\theta_o)}{\partial \theta_i \partial \theta_j} (\widehat{\theta} - \widetilde{\theta}) + R_n \end{aligned} \qquad (3.14)$$

D'autre part, en considérant le développement limité autour de la vraie valeur du paramètre θ_o, on obtient :

$$\frac{\partial}{\partial\theta}\log f(\hat{\theta}) = \frac{\partial}{\partial\theta}\log f(\theta_o) + (\hat{\theta}-\theta_o)\frac{\partial^2}{\partial\theta\partial\theta^t}\log f(\theta_o) + R'n$$

et comme $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ, on a alors :

$$\frac{\partial}{\partial\theta}\log f(\theta_o) = -(\hat{\theta}-\theta_o)\frac{\partial^2}{\partial\theta\partial\theta^t}\log f(\theta_o) + R'n$$

En remplaçant cette dernière expression dans (3.14), puis en multipliant par $f(\hat{\theta})$ et en intégrant, on obtient :

$$\begin{aligned} H(\hat{\theta}) &= -(\hat{\theta}-\tilde{\theta})\int f(\hat{\theta})\frac{\partial^2\log f(\theta_o)}{\partial\theta\partial\theta^t}(\hat{\theta}-\theta_o)dx \\ &+\frac{1}{2}(\hat{\theta}-\theta_o)^t\int f(\hat{\theta})\frac{\partial^2\log f(\theta_o)}{\partial\theta\partial\theta^t}(\hat{\theta}-\theta_o)dx \\ &-\frac{1}{2}(\tilde{\theta}-\theta_o)^t\int f(\hat{\theta})\frac{\partial^2\log f(\theta_o)}{\partial\theta\partial\theta^t}(\tilde{\theta}-\theta_o)dx + R_n \end{aligned}$$

En comme

$\int f(\hat{\theta})\frac{\partial^2\log f(\theta_o)}{\partial\theta\theta^t}dx$ est un estimateur de $\int f(\theta_o)\frac{\partial^2\log f(\theta_o)}{\partial\theta\theta^t}dx = -I_f(\theta_o)$

on en déduit que :

$$H(\hat{\theta}) = (\hat{\theta}-\theta_o)^t I_f(\theta_o)(\hat{\theta}-\tilde{\theta}) - \frac{1}{2}(\hat{\theta}-\theta_o)^t I_f(\theta_o)(\hat{\theta}-\theta_o) + \frac{1}{2}(\tilde{\theta}-\theta_o)^t I_f(\theta_o)(\tilde{\theta}-\theta_o)$$

D'autre part, en posant :

$$(\tilde{\theta}-\theta_o)^t I_f(\theta_o)(\tilde{\theta}-\theta_o) \;=\; (\tilde{\theta}-\hat{\theta}+\hat{\theta}-\theta_o)^t I_f(\theta_o)(\tilde{\theta}-\hat{\theta}+\hat{\theta}-\theta_o)$$

on aboutit à la relation :

$$H(\hat{\theta}) \approx \frac{1}{2}(\hat{\theta}-\tilde{\theta})^t I_f(\theta_o)(\hat{\theta}-\tilde{\theta})$$

Suite à un raisonnement analogue, on montre par ailleurs que :

$$G(\tilde{\theta}) \approx \frac{1}{2}(\hat{\theta}-\tilde{\theta})^t I_f(\theta_o)(\hat{\theta}-\tilde{\theta})$$

On a en définitive :

$$\hat{J}[f(\hat{\theta}),\, f(\tilde{\theta})] = H(\hat{\theta}) + G(\tilde{\theta}) \approx (\hat{\theta}-\tilde{\theta})^t I_f(\theta_o)(\hat{\theta}-\tilde{\theta})$$

$$n\hat{J}[f(\hat{\theta}),\, f(\tilde{\theta})] \approx n(\hat{\theta}-\tilde{\theta})^t I_f(\theta_o)(\hat{\theta}-\tilde{\theta}) \qquad (3.15)$$

D'où le résultat.

3.5.4 Equivalence entre le test $\hat{J}_n$ et les tests asymptotiques classiques

Nous restons toujours dans la condition de test définie par la restruction (3.4). Rappelons toutefois un résultat fondamental de la théorie asymptotique, permettant d'établir l'équivalence entre les tests W_n, R_n, *et* S_n.

Proposition

Les tests fondés sur les statistiques W_n, R_n, et S_n sont asymptotiquement équivalentes sous H_o, à la statistique notée E_n :

$$E_n = n(\hat{\theta} - \tilde{\theta})^t I(\theta_o)(\hat{\theta} - \tilde{\theta})$$

Pour montrer que la statistique $\hat{J}_n$ est équivalente à celles utilisées dans les tests classiques, il suffit de montrer qu'il en est de même avec E_n.
En utilisant la relation (3.15), on en déduit que :

$$\hat{J}_n = n\hat{J}[f(\hat{\theta}),\ f(\tilde{\theta})] \equiv E_n$$

et que $\hat{J}_n$ est donc asymptotiquement équivalente, sous l'hypothèse nulle, aux statistiques de Wald, du Multiplicateur de Lagrange et du Rapport de Vraisemblance.

Remarque :

On vient de voir - ce qui est tout à fait logique - que $\hat{J}_n$ est équivalente asymptotiquement aux statistiques usuelles.

Par contre, dans le cas des petits échantillons, ces mêmes statistiques ne sont plus nécessairement équivalentes, et peuvent donc conduire à des résultats différents. Tout l'intérêt que peut susciter $\hat{J}_n$, réside donc en ce qu'elle peut apporter à la résolution des problèmes de test, en dimension finie, par rapport aux statistiques classiques.

C'est cet aspect relatif à son comportement (en terme de puissance notamment) que nous allons traiter dans le chapitre qui va suivre.

△

Deuxième partie

Comparaison de $\widehat{J}_n$ avec les statistiques classiques et test d'adéquation fondé sur une mesure d'information

Chapitre 4

Etude comparative de $\widehat{J}$ avec les statistiques de test classiques à distance finie

---

Dans ce chapitre, nous décrivons la présentation des résultats de comparaison obtenus à partir de deux modèles, pour lesquels nous avons mis en œuvre des expériences de simulation par la méthode de Monte Carlo.

Trois sections vont composer ce chapitre. Dans la section (4.1), nous rappelons les méthodes utilisées pour interpréter les résultats obtenus. Dans la suite, nous introduisons dans la section (4.2), la comparaison des quatre statistiques de test considérés par rapport aux p-valeurs et à la fonction niveau-puissance, dans le contexte d'un modèle de régression linéaire. Enfin, dans la section (4.3), la comparaison de ces mêmes statistiques de test est faite dans le cadre d'une série d'observations issues d'un modèle exponentiel.

4.1 Méthodes d'interprétation graphique

Dans le cadre de la théorie asymptotique, la performance la pécision et l'exactitude des résultats ne se justifient que lorsque la taille de l'échantillon dont on dispose est infinie. C'est ainsi que dans le contexte de l'analyse statistique, plusieurs procédures, qui a priori peuvent sembler différentes, conduisent généralement à des

comportement asymptotiques identiques (ou équivalents).

Cependant, lorsqu'on s'intéresse à des échantillons de tailles finies, ces mêmes procédures peuvent entrainer des conclusions différentes du point de vue de l'inférence statistique. Il existe alors deux principales méthodes pour résoudre ce type de problème. La première repose sur des procédures d'approximation ou de développement asymptotiques qui engendrent très souvent des situations complexes, qui ne se prêtent pas souvent aux calculs explicites.

La deuxième approche, plus récente et dont le développement est lié à la puissance de calcul des ordinateurs, se fonde sur des résultats d'approximation obtenus à partir d'expériences de simulation (simulation par la méthode de Monte Carlo, de Metropolis ou du Boostrap).

Ces méthodes de simulation peuvent être regardées comme un outil performant qui permet soit d'approximer des quantités ou des lois de probabilité, lorsque les calculs analytiques sont compliqués voir impossibles, soit de porter un jugement en vue de la validation de résultats.

Pour mettre en évidence l'interprétation et les résultats obtenus dans la comparaison de ces différents tests, nous utilisons tout au long de ce chapitre, des méthodes dont la justification est essentiellement d'ordre graphique. Cette approche est basée principalement sur la distribution empirique des p-valeurs des statistiques de test considérées.

En particulier, dans le cas qui nous intéresse ici, la loi asymptotique des statistiques utilisées est celle de la distribution du khi-deux.

Définition

On appelle p-valeur (ou niveau de significativité marginale) associé à un test, le plus petit seuil α pour lequel on rejette l'hypothèse nulle.

Ainsi si la région critique du test considéré est définie par W_α, la p-valeur $p(x)$ est donnée par :

$$p(x) = \inf_{x \in W_\alpha} \alpha$$

Considérons, par exemple, une expérience de Monte Carlo dans laquelle N réalisations d'une statistique S sont générées par un PGD. A chacune des N répliques de la simulation, on obtient une valeur s_j $(1 \leq j \leq N)$ de S et donc une valeur p_j de la p-valeur donnée par :

$$p_j = P[S > s_j] = 1 - F_S(s_j) \tag{4.1}$$

où F_S représente la fonction de répartition asymptotique de S.

L'estimateur $\hat{F}$ de la distribution empirique F des p-valeurs, moyennes des fonctions indicatrices $\mathbf{1}_{(p_j \leq x)}$, s'écrit :

$$\hat{F}(x) \equiv \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}_{(p_j \leq x)} \tag{4.2}$$

pour chaque point x dans [0, 1].

Lorsque la distribution utilisée pour déterminer les p-valeurs p_j correspond à la loi exacte de la statistique S, on a alors, en prenant l'espérance mathématique de $\hat{F}$ dans (4.2) :

$$\begin{aligned} \mathbf{E}(\hat{F}(x)) &= P(p_j < x) = P[1 - F_S(s_j) < x] \\ &= 1 - P[1 - F_S(s_j) \geq x] = 1 - P[F_S(s_j) \leq 1 - x] \\ &= 1 - F_S(F_S^{-1}(1 - x)) = x \end{aligned}$$

Autrement dit, lorsqu'on désire étudier l'impact du niveau de signification sur la comparaison des différents tests dont on dispose, on peut chercher à tracer le graphe qui met en relief l'écart entre $F(x)$ et x, en traçant $\hat{F}(x) - x$ en fonction x.

Cette approche nous permet également de comparer très facilement les comportement des différents tests du point de vue de la puissance, en traçant les courbes de niveau-puissance, c'est-à -dire $\widehat{F^*}(x)$ en fonction de $\hat{F}(x)$ où $\hat{F}$ et $\widehat{F^*}$ représentent respectivement les estimateurs de la distribution empirique des p-valeurs sous l'hypothèse nulle et sous l'alternative.

Pour présenter les résultats obtenus par le biais d'expériences de Monte Carlo, nous nous limiterons aux deux cas suivants :

- une première section sera consacrée à l'étude de la statistique $\hat{J}_n$ par rapport à W_n, à S_n et à R_n dans le cadre précis où

on considère un modèle de régression linéaire standard ;

- nous exposerons dans une deuxième section, le cas où les observations qui constituent l'échantillon proviennent d'un modèle exponentiel.

4.2 Modèle de régression

Avant de présenter le principal critère d'évaluation du modèle, il importe de préciser un aspect fondamental qui est sous-jacent à la méthodologie que nous avons adopté : l'échantillon des observations est considéré comme provenant en général d'un processus de génération des données (PGD) d'une très "grande" complexité. Pour en étudier les points saillants et les principales caractéristiques, nous avons choisi d'élaborer pour cela, un modèle, c'est-à-dire une représentation simpliée nécessairement approximative, basée sur les données observables et sur toute autre information a priori. Se faisant, la modélisation doit donc prendre en compte à la fois la réalité intrinséque des données et les sousbassements théoriques du phénomène étudié.

En toute généralité, le modèle retenu, c'est-à-dire le modèle dont on suppose qu'il caractérise de façon adéquate les données, est d'ordre purement théorique. Néanmoins, il constitue un indicateur qui nous permet, à partir d'expressions analytiques, de procéder, dans le cas présent, à une étude comparative des propriétés du test fondé sur la statistique $\widehat{J}_n$, par rapport aux tests classiques correspondants.

Imaginons par exemple que nous ayons à effectuer la régression d'une variable y sur une variable x, à partir de n observations.
On considère alors le modèle suivant :

$$Y = X\theta + U \tag{4.3}$$

où la variable Y (supposée gaussienne), le terme d'erreur U (centré tel que les u_i de même variance) sont deux vecteurs $(n \times 1)$, le regresseur X est une matrice $(n \times k)$ et θ est un vecteur $(k \times 1)$ de paramètres inconnus.
On en déduit que la loi de U, sachant X, est la loi $N[0, \sigma^2 Id]$.

On désire tester l'hypothèse nulle $(\theta = \theta_o)$ définie comme suit :

$$\begin{cases} H_o : \ \theta = 0 \\ H_1 : \ \theta \neq 0 \end{cases}$$

On suppose en outre que $k = 1$ et, pour simplifier, $X^t = (1, \ldots, 1)$.

La résolution explicite de ce problème de test, nécessite la distinction de deux cas, suivant les quels, une solution sera ensuite proposée.

(a) Variance des perturbations connue

Lorque σ^2 est connue, le calcul de la valeur correspondant aux statistiques W_n , S_n , R_n , et $\widehat{J}_n$, engendre à la même expression.

En effet on a:

- le test de Wald qui est fondé sur l'écart $\widehat{\theta} - \theta_o$ entre l'estimateur du M.V et la vraie valeur du paramètre.

Sous l'hypothèse nulle, on a:

$$(\widehat{\theta} - \theta_o) \sim N[0, \frac{\sigma_o}{\sqrt{n}}]$$

la statistique de Wald, construite en prenant la forme quadratique associée, est alors donnée par:

$$\begin{aligned} W_n &= (\widehat{\theta} - \theta_o)^t var(\widehat{\theta})^{-1}(\widehat{\theta} - \theta_o) \\ &= n(\widehat{\theta} - \theta_o)^t \frac{1}{\sigma_o^2}(\widehat{\theta} - \theta_o) \end{aligned}$$

- la statistique du score qui est obtenue à partir de l'expression

$$\widehat{\lambda} = -\frac{2n}{\sigma_o^2}(\theta_o - \widehat{\theta})$$

ainsi, sous l'hypothèse H_o, on peut écrire:

$$\widehat{\lambda} \sim N[0, \frac{4n}{\sigma_o^2}]$$

la statistique du score prend alors la forme suivante:

$$S_n = n(\widehat{\theta} - \theta_o)^t \frac{1}{\sigma_o^2}(\widehat{\theta} - \theta_o)$$

- la différence des valeurs de la Log-vraisemblance montre que

$$\mathcal{L}(\widehat{\theta}) - \mathcal{L}(\theta_o) = \frac{-1}{2\sigma_o^2}(||Y - X\widehat{\theta}||^2 - ||Y - X\theta_o||^2)$$

d'après le théorème de Pythagore, cette expression peut se mettre sous la forme:

$$\mathcal{L}(\widehat{\theta}) - \mathcal{L}(\theta_o) = \frac{1}{2\sigma_o^2}||X\widehat{\theta} - X\theta_o||^2$$

et on obtient en fin compte :

$$R_n = \frac{1}{\sigma_o^2}||X\widehat{\theta} - X\theta_o||^2 = n(\widehat{\theta} - \theta_o)^t \frac{1}{\sigma_o^2}(\widehat{\theta} - \theta_o)$$

- l'estimateur de la divergence $\widehat{J}_n$, est obtenu après un calcul sans difficulté :

$$\widehat{J}_n \quad = n(\widehat{\theta} - \theta_o)^t \frac{1}{\sigma_o^2}(\widehat{\theta} - \theta_o)$$

où l'estimateur du maximum de vraisemblance non contraint de θ_o coïncide avec l'estimateur des moindres carrés ordinaires, à savoir :

$$\widehat{\theta} = (X^t X)^{-1} X^t Y$$

On en déduit donc que ces statistiques de test, qui sont asymptotiquement équivalentes, sont égales à distance finie, et lorsque la variance des perturbations est connue, à la même expression :

$$n\widehat{J} \equiv W_n \equiv S_n \equiv R_n = n(\widehat{\theta} - \theta_o)^t \frac{1}{\sigma_o^2}(\widehat{\theta} - \theta_o)$$

Une conséquence sous-jacente à ce résultat, est que lorsque la taille de l'échantillon est fixée, les propriétés issues de ces différentes statistiques de test sont parfaitement identiques.

(b) Variance des perturbations inconnue

Lorsque σ^2 est inconnue, il faudra alors l'estimer par $\widehat{\sigma}^2$, donnée par :

$$\widehat{\sigma}^2 = \frac{1}{n}||Y - \widehat{\theta}X||^2$$

Sous l'hypothèse nulle $\theta_o = 0$ (ie $Y = U$), nous allons considérer l'estimateur suivant :

$$\widehat{\sigma}_o^2 = \frac{1}{n}||Y||^2 = \frac{1}{n}Y^t Y$$

Lorsque la taille n de l'échantillon est fixe, les trois tests classiques prennent les formes simples suivantes :

$$S_n \quad = \quad \frac{nF}{n - 1 + F}$$

$$\begin{aligned} W_n &= \frac{n}{n-1}F \\ R_n &= n\log(1+\frac{F}{n-1}) \\ F &= \frac{(n^{-1/2}X^tY)^2}{Y^tY - n^{-1/2}X^tY}(n-1) \end{aligned}$$

où F représente la statistique de Fisher.

Un calcul rapide montre par ailleurs que la statistique $\widehat{J}_n$ peut se mettre sous la forme d'une expression, également fonction de F. On a en effet :

$$\widehat{J}_n = \frac{n}{2}(\frac{\widehat{\sigma}^2}{\widehat{\sigma}_o^2} + \frac{\widehat{\sigma}_o^2}{\widehat{\sigma}^2}) + \frac{n}{2}(\widehat{\theta} - \widehat{\theta}_o)^2(\frac{1}{\widehat{\sigma}_o^2} + \frac{1}{\widehat{\sigma}^2}) - n$$

En utilisant les estimateurs $\widehat{\sigma}^2$ et $\widehat{\sigma}_o^2$ définis plus haut, on aboutit à une relation de proportionnalité avec F, donnée par :

$$\widehat{J}_n = \frac{n}{n-1}F$$

Dans ce contexte, toutes ces statistiques suivent asymptotiquement, sous H_o, une loi du χ^2. On remarquera, dans le cas particulier considéré ici, que les quantiles de leur loi exacte sont facilement calculables puisqu'elles s'écrivent toutes comme des fonctions croissantes de la statistique de Fisher. On en déduit que dans le cas d'un modèle linéaire, lorsque la taille des échantillons est finie, la famille des lois de Fisher F (ce qui est conforme à l'approche classique) permet de résoudre le problème de test défini ci-dessus.

4.2.1 Présentation des résultats

Nous présentons ici les résultats issus d'une suite d'expériences de Monte Carlo, à partir du modèle de régression linéaire standard défini en (4.3). Nous générons d'abord une série u_t à partir de la loi de Laplace-Gauss $N[0,\ \sigma]$ avec $\sigma = 1$ par exemple.

Ensuite y_t est engendrée sur la base de 5.000 répliques de simulation par Monte Carlo obtenues à partir des PGD (4.2) et calculées sous l'hypothèse nulle ($H_o :\ \theta_o = 0$).

On sait en outre, que si la distribution de la statistique utilisée pour calculer les p-valeurs p_j est exacte, alors on a nécessairement : $\widehat{F}(x) \equiv x$.

Ainsi, pour une taille d'échantillon $n = 20$, les graphes de l'écart entre $\widehat{F}(x)$ et x en fonction des x, permettent une interprétation simple des p-valeurs obtenues à partir des quatres statistiques étudiées par rapport au niveau nominal.

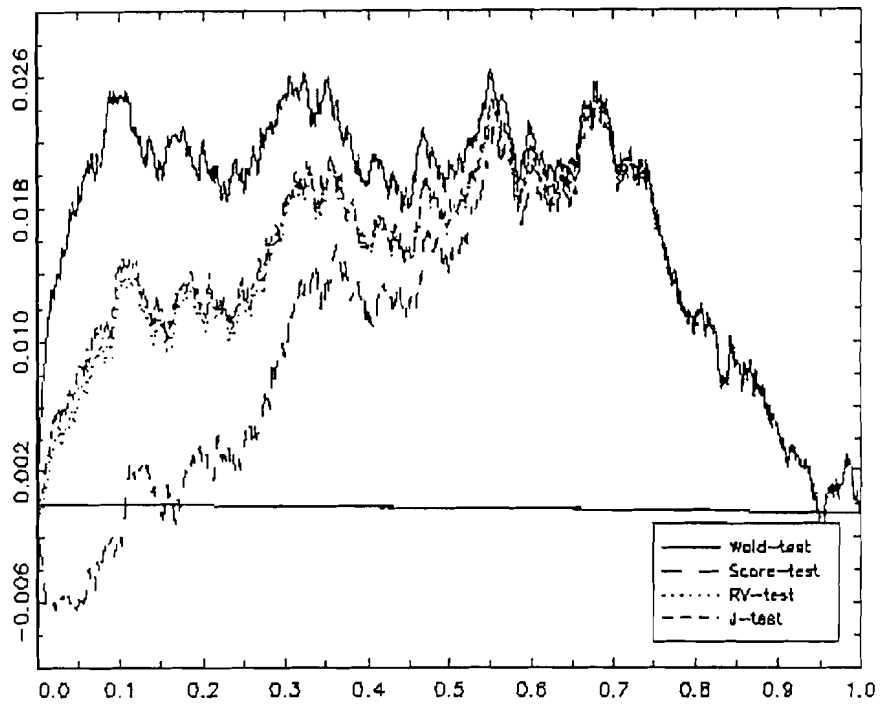

Figure 1: Courbes de $\widehat{F}(x) - x$ en fonction x

Dans la Figure 1, les résultats obtenus pour une taille d'échantillon égale à $n = 20$, montrent que la statistique du score S_n commence par sous-rejeter le test au voisinage de [0, 0.1], ensuite elle sur-rejette le test partout ailleurs. Les trois autres tests sur-rejettent systématiquement l'hypothèse nulle sur l'intervalle [0, 1]. Cette méthode, qui montre que ces statistiques de test ont un comportement tout à fait acceptable, vu sous l'angle des p-valeurs, ne permet visiblement pas de disposer d'un pouvoir de discrimination performant.

Lorsqu'on désire maintenant mettre en compétition les quatre statistiques de test considérées ici, du point de vue de leur performance en terme de puissance, on peut représenter les résultats sous forme de courbes de niveau-puissance. Il suffit en effet d'utiliser une méthode graphique fondée sur l'estimateur de la fonction de distribution empirique des p-valeurs $\widehat{F^*}$ obtenue à partir d'une série d'expériences

de Monte Carlo, dans laquelle les données sont générées par le PGD (4.3) sous l'hypothèse alternative.

Les courbes de niveau puissance sont alors obtenues en traçant $\widehat{F^*}$ en fonction de l'estimateur de la fonction de distribution empirique $\widehat{F}$.

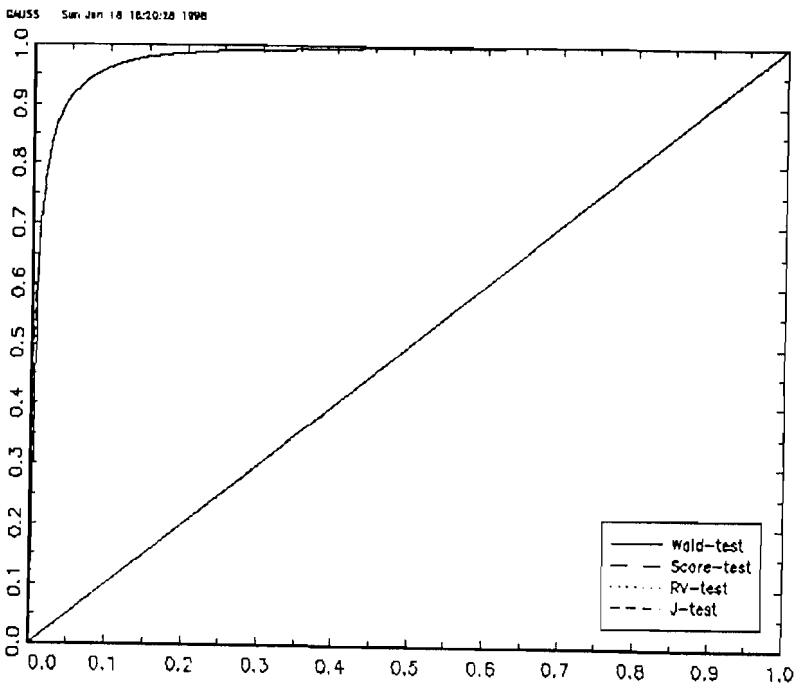

Fig2: Courbes de niveau-puissance pour $\theta_o = 0$ et $\theta = 0.8$

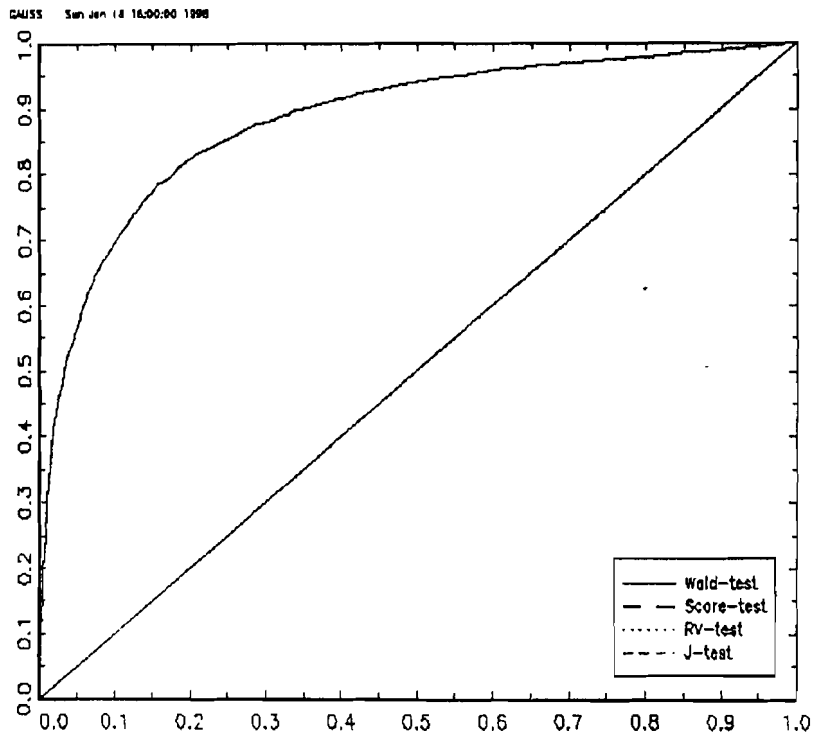

Fig3: Courbes de niveau-puissance pour $\theta_o = 0$ et $\theta = 0.5$

Comme l'on pouvait le prévoir, les courbes de niveau-puissance relatives aux quatre statistiques de test sont parfaitement identiques. Cela s'explique par le fait que, sous l'hypothèse nulle, toutes ces statistiques peuvent s'écrire comme fonction croissante, de la statistique de Fisher.

Les figures 2 et 3 décrivent le comportement de ces courbes de niveau-puissance lorsque θ varie. Evidemment, il y a une augmentation de la puissance lorsque θ s'éloigne de sa vraie valeur $\theta_o = 0$.

Regardons maintenant, un cas plus intéressant, où les observations suivent une loi exponentielle.

4.3 Cas de la loi exponentielle

Nous supposons ici que les observations sont issues d'une loi exponentielle de paramètre β. Cette loi est l'analogue continue de la loi géométrique. En effet sa propriété d' absence de postaction

$$Prob[X > t + s/X > s] = Prob[X > t]$$

la rend essentielle en théorie des processus markoviens discontinus, et elle est très appréciée pour l'analyse d'évènements en relation avec les temps d'arrêt, la durée du chômage etc ...

Ainsi si X est la variable aléatoire qui résume l'échantillon considéré, sa densité est alors donnée par :

$$X \sim f(x, \beta) = \begin{cases} \frac{1}{\beta} exp\{-\frac{x}{\beta}\} \\ \beta > 0 \; ; \quad x \geq 0 \end{cases} \tag{4.4}$$

Considérons le problème de test défini par les hypothèses suivantes :

$$\begin{cases} H_o : \quad \beta = \beta_o \\ H_1 : \beta \neq \beta_o \end{cases} \tag{4.5}$$

L'estimateur du maximum de vraisemblance, associé au paramètre β, s'obtient au travers de l'expression :

$$\widehat{\beta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

On souhaite maintenant, dans le cadre du modèle exponentiel défini en (4.4), d'étudier et de comparer les statistiques de test asymptotiques usuelles avec celle fondée sur la divergence J.

On obtient alors les expressions ci-dessous, suivant les statistiques considérées :

- **Test de Wald :**

La statistique de Wald associée aux hypothèses (4.5), d'après le modèle (4.4), peut s'écrire sous H_o, suivant la formulation :

$$W_n \quad = \quad (\widehat{\beta}_n - \beta_o)^t \; var^{-1}(\widehat{\beta}_n) \; (\widehat{\beta}_n - \beta_o)$$

$$= n(\widehat{\beta}_n - \beta_o)^t \beta_o^{-2} (\widehat{\beta}_n - \beta_o) \tag{4.6}$$

$\widehat{\beta}_n$ étant l'estimateur du maximum de vraissemblance.

- **Test du Rapport de vraisemblance**:

Dans les mêmes conditions définies en (4.4) et (4.5), la formulation de la statistique du rapport de vraisemblance, sous l'hypothèse nulle, est donnée par la relation:

$$R_n = 2n[\log \frac{\beta_o}{\widehat{\beta}_n} + \frac{\widehat{\beta}_n}{\beta_o} - 1] \tag{4.7}$$

- **Test fondé sur la divergence:**

L'expression de la valeur de la statistique $\widehat{J}_n$, sous les conditions (4.4) et (4.5), prend une forme relativement simple, toujours sous l'hypothèse nulle:

$$\widehat{J}_n = (\widehat{\beta}_n - \beta_o)^t (\widehat{\beta}_n \beta_o)^{-1} (\widehat{\beta}_n - \beta_o) \tag{4.8}$$

Maintenant que ces statistiques sont parfaitement spécifiées, dans le cadre du modèle retenu, on peut alors chercher à préciser, dans le paragraphe qui suit, certaines propriétés de la statistique $\widehat{J}_n$.

4.3.1 Propriétés de puissance de ces différentes statistiques

Dans cette section, nous utilisons des développements par approximation pour examiner les propriétés théoriques liées aux statistiques de Wald, du Rapport de vraisemblance et celle de la divergence J.

Considérons les développement à l'ordre 2, sous l'hypothèse alternative de ces estimateurs

($\beta \neq \beta_o$). Dans ce contexte, on a: $\widehat{\beta}_n \longrightarrow \beta_1 \neq \beta_o$.

On obtient:

1) Pour la statistique de Wald

$$\begin{aligned} W_n &= n(\widehat{\beta}_n - \beta_o)^t \frac{1}{\beta_o^2} (\widehat{\beta}_n - \beta_o) \\ &= n(\widehat{\beta}_n - \beta_1)^t \frac{1}{\beta_o^2} (\widehat{\beta}_n - \beta_1) + n(\beta_1 - \beta_o)^t \frac{1}{\beta_o^2} (\beta_1 - \beta_o) \end{aligned}$$

$$+2n(\hat{\beta}_n - \beta_1)^t \frac{1}{\beta_o^2}(\beta_1 - \beta_o)$$

Posons :

$$A_o = (\beta_1 - \beta_o)^t \frac{1}{\beta_o^2}(\beta_1 - \beta_o)$$

$$A_1 = (\hat{\beta}_n - \beta_1)^t \frac{1}{\beta_o^2}(\hat{\beta}_n - \beta_1)$$

$$A_2 = (\hat{\beta}_n - \beta_1)^t \frac{1}{\beta_o^2}(\beta_1 - \beta_o)$$

Soit alors :

$$W_n \simeq nA_o + n[A_1 + 2A_2]$$

2) Pour la statistique du Rapport de Vraisemblance

$$\begin{aligned}\ell(\hat{\beta}_n) &= \ell(\beta_o) + (\hat{\beta}_n - \beta_o)\frac{\partial}{\partial\beta}\ell(\beta_o) + \frac{1}{2}(\hat{\beta}_n - \beta_o)^2\frac{\partial^2}{\partial\beta^2}\ell(\beta_o) + R \\ &= \ell(\beta_o) + (\hat{\beta}_n - \beta_o)(\frac{-n}{\beta_o} + \frac{-n\hat{\beta}_n}{\beta_o^2}) + \frac{1}{2}(\hat{\beta}_n - \beta_o)^2(\frac{n}{\beta_o^2} - \frac{-2n\hat{\beta}_n}{\beta_o^3}) + R\end{aligned}$$

Autrement dit :

$$\begin{aligned}\ell(\hat{\beta}_n) - \ell(\beta_o) &= \frac{n}{2}\frac{(\hat{\beta}_n - \beta_o)^2}{\beta_o^2} + (\hat{\beta}_n - \beta_o)(\frac{n\hat{\beta}_n - n\beta_o}{\beta_o^2}) - \frac{n\hat{\beta}_n}{\beta_o^3}(\hat{\beta}_n - \beta_o)^2 + R \\ &= \frac{n}{2}\frac{(\hat{\beta}_n - \beta_o)^2}{\beta_o^2} + (\frac{n(\hat{\beta}_n - \beta_o)^2)}{\beta_o^2}) - \frac{n\hat{\beta}_n(\hat{\beta}_n - \beta_o)^2)}{\beta_o^3} + R \\ &= \frac{n}{2}\frac{(\hat{\beta}_n - \beta_o)^2}{\beta_o^2} + (\frac{n(\hat{\beta}_n - \beta_o)^2)}{\beta_o^3})(\beta_o - \hat{\beta}_n) + R\end{aligned}$$

On obtient enfin de compte :

$$R_n \simeq W_n + 2n\frac{(\beta_o - \hat{\beta}_n)^3}{\beta_o^3}$$

Ce qui entraine :

$$R_n \simeq nA_o + n[A_1 + 2A_2 + 2A_3]$$

avec

$$A_3 = \frac{(\beta_o - \hat{\beta}_n)^3}{\beta_o^3}$$

3) Pour la statistique de la divergence

Posons

$$\hat{J}_n = n[\int f(x/\hat{\beta}_n)\log\frac{f(x/\hat{\beta}_n)}{f(x/\beta_o)}dx + \int f(x/\beta_o)\log\frac{f(x/\beta_o)}{f(x/\hat{\beta}_n)}dx]$$

$$= n[K(\widehat{\beta}_n) + L(\widehat{\beta}_n)]$$

avec

$$K(\widehat{\beta}_n) = \int f(x/\widehat{\beta}_n \log \frac{f(x/\widehat{\beta}_n)}{f(x/\beta_o)} dx \qquad et \qquad L(\widehat{\beta}_n) = \int f(x/\beta_o) \log \frac{f(x/\beta_o)}{f(x/\widehat{\beta}_n)} dx$$

$$\begin{aligned} \widehat{J}_n(\widehat{\beta}_n) &\simeq \frac{n(\widehat{\beta}_n - \beta_o)^2}{\beta_o^2} - \frac{n(\widehat{\beta}_n - \beta_o)^3}{\beta_o^3} \\ &\simeq W_n - \frac{n(\widehat{\beta}_n - \beta_o)^3}{\beta_o^3} \end{aligned}$$

Ce qui donne, d'après le même raisonnement que précédemment :

$$\widehat{J}_n \simeq nA_o + n[A_1 + 2A_2 + A_3]$$

Soit :

$$\widehat{J}_n \simeq \frac{1}{2}[W_n + R_n] \tag{4.9}$$

La relation précédente (4.9) montre que ces statistiques de test ne sont pas égales en dimension finie. On en conclut que sous l'hypothèse alternative, l'estimateur de la divergence $\widehat{J}_n$ apparaît comme une *"moyenne"* des statistiques de Wald et du rapport de vraisemblance et que par conséquent, leur puissance ne peut être égale, notamment si l'on s'intéresse à des échantillons de tailles finies.

Pour des raisons pratiques, on peut procéder à une évaluation locale de la puissance du test généré en fonction de $\widehat{J}_n$ et la comparer ensuite, avec celles relevant des statistiques de W_n et R_n, afin d'obtenir une confirmation de leur inégalité, dans le cas présent.

En effet sous l'alternative $H_1 : \beta \neq \beta_o$, on considère que la loi limite de $\widehat{J}_n$ est une loi du khi-deux ($\chi^2(d, \lambda^2)$), avec un paramètre de non-centralité fourni par :

$$\lambda^2 = n\frac{(\beta - \beta_o)^2}{\beta\beta_o}$$

Pour obtenir une approximation de la distribution du χ^2-décentrée, par une loi du χ^2 centrée, on pose alors :

$$\chi^2(d, \lambda^2) \simeq \gamma\chi^2(c, 0)$$

$$\text{où} \quad \gamma = \frac{d+2\lambda^2}{d+\lambda^2} \quad \text{et} \quad c = \frac{(d+\lambda^2)^2}{d+2\lambda^2}$$

En conséquence, pour une valeur c_α donnée, la puissance approximative P de $\hat{J}_n$ s'obtient aisément, par le biais de l'expression définie par :

$$P = Prob[\hat{J}_n > c_\alpha] \simeq Prob[\chi^2\Big(\frac{(d+\lambda^2)^2}{d+2\lambda^2}\Big) > c_\alpha \frac{d+\lambda^2}{d+2\lambda^2}] \qquad (4.10)$$

Cette relation (4.10) nous fournit une expression analytique qui permet d'évaluer l'impact du paramètre de non centralité λ^2 sur la puissance de $\hat{J}_n$.

Par ailleurs, cette même relation, ainsi que celles correspondant à W_n et R_n, peuvent permettre, ne serait-ce que localement une comparaison de la puissance, ici, à partir de deux exemples ou β sera respectivement supérieur et inférieur à β_o.

4.3.2 Exemple d'illustration

Reprenons l'exemple du modèle exponentiel (4.4) défini précédemment.

On désire tester l'hypothèse nulle

$$\begin{cases} H_o : \ \beta = \beta_o \\ H_1 : \ \beta \neq \beta_o \end{cases} \qquad (4.11)$$

On suppose toujours que $\hat{\beta}_n$ est l'estimateur du maximun de vraisemblance de β.

a) La valeur du paramètre β est supérieure à β_o ($\beta > \beta_o$)

Considérons les paramètres de non-centralité μ^2 , η^2, et η^2, correspondant respectivement à W_n, $n\hat{J}$, R_n et supposons en outre la donnée des valeurs suivantes : $\beta = 1.5$; $\beta_o = 1$; $n = 50$ et $\alpha = 0.05$.

Ce qui entraine :

$$\mu^2 = n\frac{(\beta-\beta_o)^2}{(\beta)^2} = 5.56$$

$$\eta^2 = n\frac{(\beta-\beta_o)^2}{\beta\beta_o} = 8.33$$

$$r^2 = 2n[\log\frac{\beta_o}{\beta} + \beta - 1] = 9.45$$

On en déduit alors la puissance des trois tests considérés ci-dessus :

- *Puissance du test de Wald*

$$\begin{aligned} P_w &= Prob[W_n > 3.84] \\ &= Prob[\chi^2(3.548) > 2.078] \simeq 0.57 \end{aligned}$$

- *Puissance de la divergence $\widehat{J}_n$*

$$\begin{aligned} P_j &= Prob[\widehat{J}_n > 3.84] \\ &= Prob[\chi^2(4.930) > 2.0287] \simeq 0.85 \end{aligned}$$

- *Puissance du test du rapport de vraisemblance*

$$\begin{aligned} P_r &= Prob[R_n > 3.84] \\ &= Prob[\chi^2(5.4893) > 2.0164] \simeq 0.92 \end{aligned}$$

b) La valeur du paramètre β est inférieure à β_o ($\beta < \beta_o$)

On considère ici, le cas ou : $\beta = 0.6$; $\beta_o = 1$; $n = 50$ et $\alpha = 5\%$.
On obtient dans ce cas :

$$\mu^2 = n\frac{(\beta - \beta_o)^2}{(\beta)^2} = 22.22$$

$$\eta^2 = n\frac{(\beta - \beta_o)^2}{\beta\beta_o} = 13.33$$

$$r^2 = 2n[\log\frac{\beta_o}{\beta} + \beta - 1] = 11.08$$

On peut alors évaluer, dans ce cas, la puissance correspondant à chaque statistique considérée.
D'où l'on a :

- *Puissance du test de Wald*

$$\begin{aligned} P_w &= Prob[W_n > 3.84] \\ &= Prob[\chi^2(11.8667) > 1.9623] \simeq 0.99 \end{aligned}$$

- *Puissance du test fondé sur la divergence $\widehat{J}_n$*

$$\begin{aligned} P_j &= Prob[\widehat{J}_n > 3.84] \\ &= Prob[\chi^2(7.4257) > 1.9894] \simeq 0.98 \end{aligned}$$

- *Puissance du test du rapport de vraisemblance*

$$\begin{aligned} P_r &= Prob[R_n > 3.84] \\ &= Prob[\chi^2(6.3021) > 2.0029] \simeq 0.96 \end{aligned}$$

On peut conclure, d'après **a**) et **b**) que les puissances calculées suivant les procédures de test utilisant les statistiques W_n , $\widehat{J}_n$ et R_n, prennent lorsque l'on considère de petits échantillons (ici $n = 50$), des valeurs différentes selon que le paramètre β soit supérieur ou inférieur à β_o. C'est ainsi qu'une analyse de ces résultats montre que lorsque $\beta > \beta_o$, une comparaison de ces puissances, au niveau 5%, entraine: $P_w < P_j < P_r$.

Et lorsque $\beta < \beta_o$, on aboutit, pour un même niveau de signification, au résultat inverse: $P_r < P_j < P_w$. Ce qui montre comme l'on pouvait s'en douter, qu'il n'existe pas de test uniformément plus puissant que tous les autres.

Afin d'étendre cette comparaison locale ($\alpha = 5\%$) au cas d'une étude comparative uniforme ($\alpha \in [0, 1]$), on propose de procéder par la méthode graphique soulignée ci-dessus. C'est précisément l'objet du paragraphe qui va suivre où notre centre d'interêt est de savoir comment se comportent en pratique ces différentes statistiques de test en dimension finie.

4.3.3 Analyse des courbes de puissance par une méthode graphique

La comparaison en terme de puissance de ces tests, repose sur le fait que suivant la valeur prise par le paramètre β, par rapport à l'hypothèse nulle β_o, on obtient des valeurs différentes pour les statistiques considérées.

Pour étudier le comportement de puissance relative à ces tests, nous allons, procéder par une méthode fréquemment utilisée en statistique, qui consiste regarder l'impact de cette propriété, lorsqu'on s'écarte un peu de l'hypothèse à tester.

Nous allons ensuite comparer les trois tests (W_n, R_n et $\widehat{J}_n$), par le biais des courbes de niveau-puissance. Pour cela, nous introduisons un modèle dans lequel, l'échantillon des observations est considéré comme provenant d'un processus de génération des données.

Nous poserons :

$$\begin{cases} Y \sim \dfrac{1}{\beta}\, exp\{-\dfrac{x}{\beta}\} \\ x > 0\,; \quad \beta = \beta_o(1+\epsilon); \quad \beta_o > 0 \quad \epsilon \in]-1,\ +\infty[\end{cases} \tag{4.12}$$

où β_o désigne la vraie valeur du paramètre (cf à l'hypothèse nulle) et ϵ un réel donné.

Ainsi, on obtient des valeurs correspondantes du paramètre, qui sont supérieures ou inférieures à β_o suivant que ϵ soit positif ou négatif, la valeur $\epsilon = 0$ redonnant l'hypothèse nulle.

Les calculs de la simulation sont basés sur 5000 répliques d'expériences de Monte-Carlo, à partir du modèle (4.12).

La figure 4 montre les courbes de niveau-puissance des trois tests considérés pour le cas $n = 50$ et $\epsilon = 0.4$.

L'interprétation évidente de ces courbes, montrent que pour une cinquantaine d'observations et pour $\epsilon = 0.4$, le test du rapport de vraisemblance R_n est plus puissant que celui fondé sur la statistique $\widehat{J}_n$. Le test de Wald, quant à lui, reste moins performant que ces deux concurrents.

L'interprétation de la figure 5, conclut à un résultat inverse, lorsque l'on choisit une valeur négative pour ϵ (ici on a considéré $\epsilon = -0.3$ pour une même taille d'échantillon), en ce sens que la puissance du test associé à la statistique de Wald est supérieure à celle générée par la statistique de la divergence $\widehat{J}_n$, qui à son tour fait mieux que celle issue de la statistique du rapport de vraisemblance. Ce qui confirme ainsi les résultats obtenus par approximation dans l'exemple (4.3.2).

La Figure 6 permet de constater, conformément à la logique de la théorie des tests, que la performance obtenue en fonction de $\widehat{J}_n$, dimunie fortement en terme

de puissance lorsque ϵ tend vers 0 (hypothèse nulle), par valeurs positives (Figure 6a : $\epsilon = 0.2;\ 0.4;\ 0.6$) ou négatives (Figure 6b : $\epsilon = -0.4;\ -0.3;\ -0.2$).

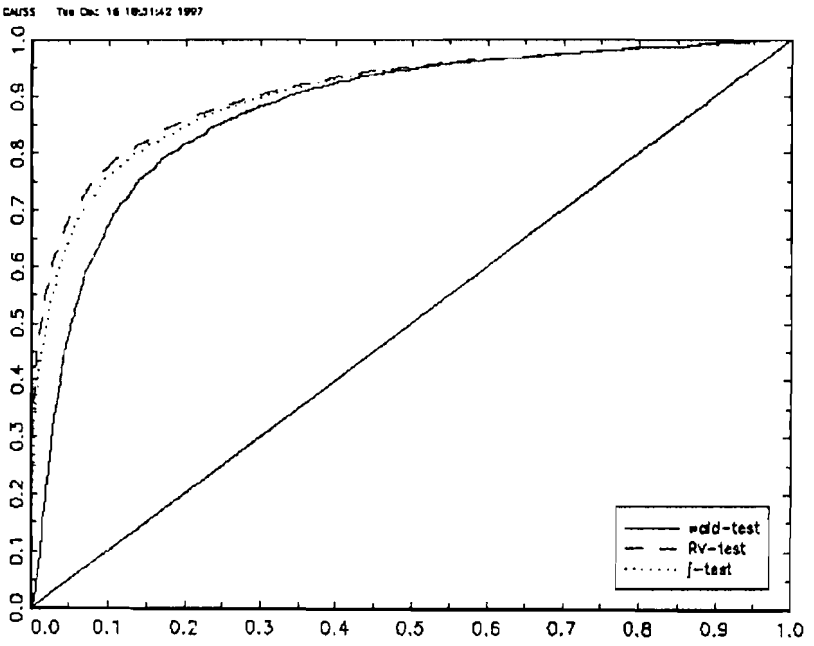

Fig 4: Courbes niveau-puissance pour $n = 50$ et $\epsilon = 0.4$

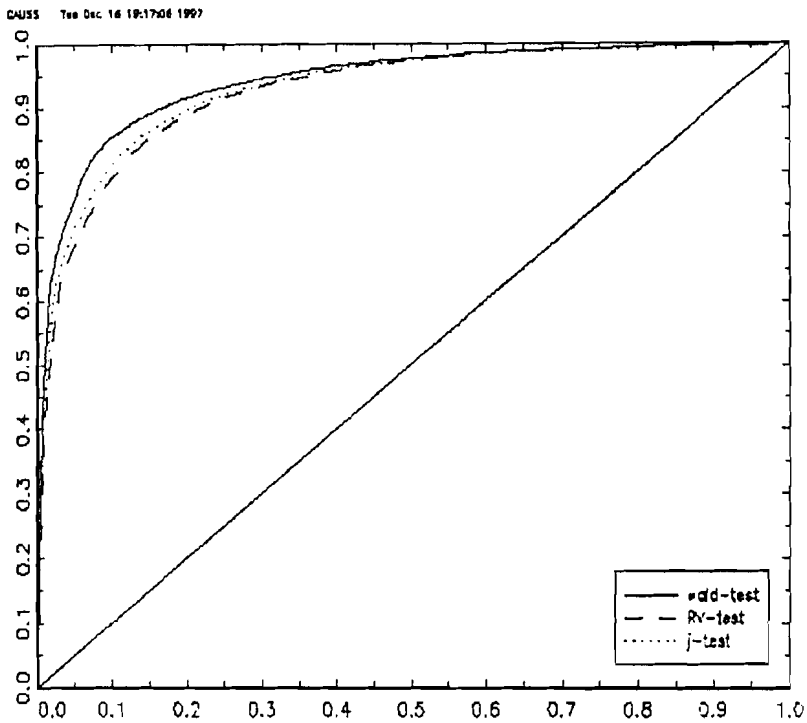

Fig 5: Courbes niveau-puissance pour $n = 50$ et $\epsilon = -0.3$

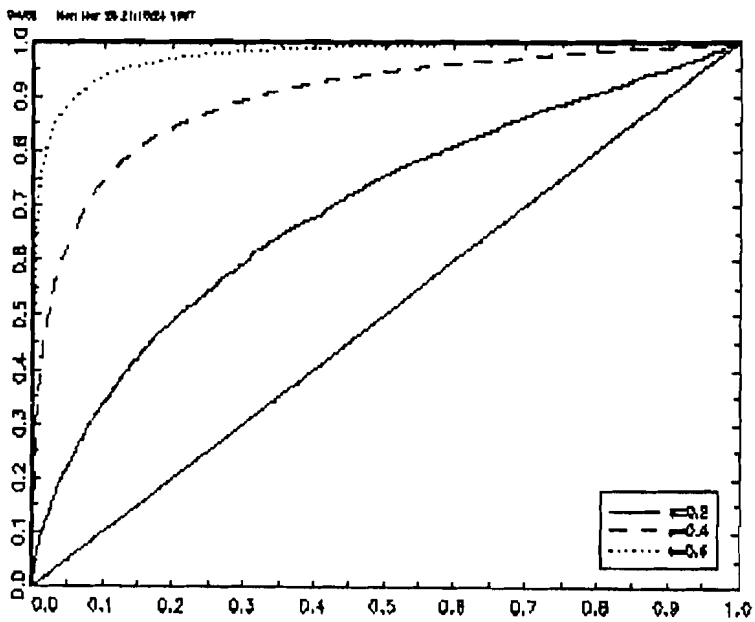

Fig 6a: Courbes niveau-puissance pour $n = 50$ et $\epsilon = 0.2,\ 0.4;\ 0.6$

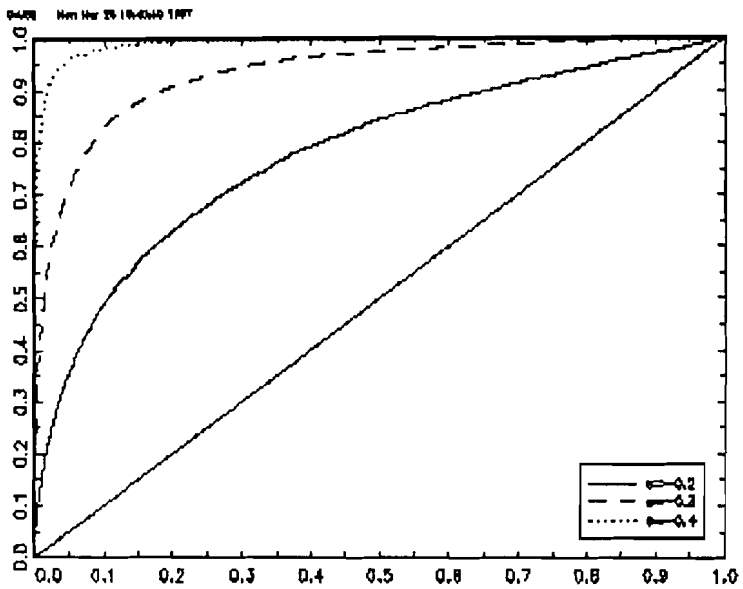

Fig 6 b: Courbes niveau-puissance pour $n = 50$ et $\epsilon = -0.4;\ -0.3;\ -0.2$

En résumé, nous avons donc considérer une mesure J, appartenant à la famille des critères de divergence de type Kullback, afin de proposer une statistique de test $\widehat{J}_n$, que l'on compare ensuite avec les statistiques de test classiques.

L'avantage majeur associé à $\widehat{J}_n$ est que d'une part, elle présente une propriété de symétrie, et que d'autre part, elle suit, sous l'hypothèse nulle - comme c'est le cas des statistiques de test habituelles - une loi du khi-deux, précisément lorsque la taille de l'échantillon tend vers l'infini.

L'intérêt de $\widehat{J}_n$ est mis en évidence, lorsque l'on se restreint à des échantillons finis. Nous avons donc considérés ici, des exemples simples (la méthode pouvant se généraliser à des cas plus ou moins complexes), pour montrer que grâce à la symétrie $\widehat{J}_n$ possède, par rapport aux statistiques de Wald et du Rapport de vraisemblance, une propriété de robustesse, lorsque l'on considère certains types de lois.

Chapitre 5

Test d'ajustement et test de sélection à partir d'une mesure de divergence

---

5.1 Introduction

Les tests d'adéquation d'un modèle à un échantillon se fondent en général sur des statistiques dont la distribution asymptotique suit une loi du Khi-deux. De nombreux auteurs parmi lesquels Watson (1959), Moore (1978,1986) ont étudiés ces types de statistiques qui peuvent souvent s'écrire sous formes quadratiques. C'est ainsi que la statistique d'ajustement la plus fréquemment utilisée est celle de Pearson, du fait qu'elle fournit une mesure naturelle de la divergence évaluée par l'écart entre la fréquence empirique et la probabilité théorique. La méthode proposée par Pearson consiste à regrouper les données dont on dispose en M classes et à calculer la "*distance du khi-deux*" définie par :

$$Q_n = n \sum_{i=1}^{M} \frac{(\widehat{p}_i - t_i)^2}{t_i}$$

où n est la taille de l'échantillon considéré, $\widehat{p}_i$ et t_i désignent respectivement la proportion empirique et la fréquence théorique de la classe i correspondante.

Le foisonnement de la diversité des domaines d'applications du test d'adéquation fondé sur la statistique du khi-deux, notamment dans les sciences sociales et en économétrie entre autres, nous incitent à rechercher de nouvelles statistiques en vue de constuire des tests d'ajustement avec un meillleur comportement par rapport à la puissance.

Nous nous proposons dans les pages qui suivent de discuter quelques mesures de divergence autour desquelles nous établirons un test d'adéquation dans le cadre d'un modèle paramétrique. Ce test sera fondé sur un estimateur $\widehat{\Delta}_r$ obtenu à partir d'une classe de mesures de divergence Δ_r. Une étude sera ensuite menée pour déterminer la valeur du paramètre r pour laquelle, le test devient optimal en terme de puissance et nous le comparerons ensuite avec le test classique de Pearson ainsi que celui de Kolmogorov-Smirnov.

Nous tenterons, chemin faisant, d'en donner une application à la résolution d'un problème de test de choix entre deux distributions, en s'appuyant sur le critère d'Akaike (A.I.C-1973). Toutefois une difficulté bien connue, liée à l'utilisation directe de ce critère de choix, est qu'il ne précise pas le seuil de confiance que l'on peut accorder au modèle retenu.

Pour tenir compte du niveau de signification inhérente à toute décision statistique, Vuong et Wang (1993) proposent l'usage d'un test asymptotiquement normal lorsque la sélection de modèle est fondée sur des statistiques de type Pearson.

De façon analogue à l'approche de Vuong et Wang, nous suggérons, dans une deuxième partie, d'appuyer le problème du test de choix entre deux modèles paramétriques sur une statistique construite à partir de la distance $\Delta_{1/2}$ et nous établissons une comparaison, selon la loi suivie par les observations, entre les deux modèles.

5.2 Estimateur de la mesure Δ_r

5.2.1 Définitions et hypothèses

La notion de distance entre deux distributions présente un intérêt considérable en théorie de l'information plus particulièrement en analyse statistique. Dans cette section nous présentons les hypothèses de base du modèle que nous allons considérer, les estimateurs des paramètres d'intérêt ainsi que la statistique fondée sur la mesure de divergence Δ_r qui formalisent la procédure de test.

La première mise en œuvre d'une généralisation au travers d'un paramètre scalaire, a été introduite par Rényi en 1961, en posant:

$$\begin{cases} D_r^1[P, Q] = (r-1)^{-1} ln\{\sum_{i=1}^{n} p_i^r q_i^{1-r}\} \\ r \neq 1, \quad r > 0 \end{cases} \tag{5.1}$$

comme mesure de proximité entre deux distributions données P et Q.
Un peu plus tard, Sharma et Mittal en (1977) ont mené une étude de généralisation basée sur deux paramètres, incluant la mesure de Rényi comme cas limite, en posant :

$$\begin{cases} D_r^s[P, Q] = (s-1)^{-1}\{\{\sum_{i=1}^{n} p_i^r q_i^{1-r}\}^{\frac{s-1}{r-1}} - 1\} \\ r \neq 1, s \neq 1 \quad r > 0 \quad s > 0 \end{cases} \tag{5.2}$$

La mesure d'information D_r^r correspondant au cas $r = s$, a été largement étudiée par plusieurs auteurs. Pour des développement plus généraux, on pourra se referer aux contributions - entre autres - de Mathaï et Rathie (1975) et Tanéja (1979).

En vue de présenter une mesure symétrique déduite de $D_r^r[P, Q]$, nous pouvons envisager un critère d'évaluation de la proximité entre les distributions P et Q, en considérant la divergence $\Delta_r[P, Q]$ définie par:

$$\Delta_r[P, Q] = D_r^r[P, Q] + D_r^r[P, Q] \quad ; \quad \Delta_1[P, Q] = \lim_{r \longrightarrow 1} \Delta_r[P, Q] \tag{5.3}$$

C'est autour de cette mesure $\Delta_r[P, Q]$, ainsi que son estimateur $\hat{\Delta}_r[P, Q]$, que nous organisons une méthode de test d'ajustement d'une série d'observations à un modèle paramétrique donné.

Toutefois, pour mettre en œuvre cette procédure de test, des conditions principales doivent être observées sous forme d'hypothèses:

Hypothèse 1

On suppose que les observations X_i, $i = 1,2,\ldots$ sont indépendantes et identiques avec une distribution commune H.

L'espace d'échantillonnage Ξ est partionné en M classes deux à deux disjointes: $E_1,E_2,\ldots,E_M$.

Considérons un modèle $H_\theta = \{H(x,\theta)\ ;\ x \in \Xi, \theta \in \Theta \subset \mathbf{R^k}\}$ et le vecteur des probabilités associées à la répartition:

$$h(\theta) = (h_1(\theta),h_2(\theta),\ldots,h_M(\theta))$$

avec:

$$h_i(\theta) = \int_{E_i} dH(x,\theta) \quad i = 1,2,\ldots,M \tag{5.4}$$

Hypothèse 2

On suppose que $h_i(\theta)$ vérifie les conditions de régularité suivantes:

- (i) le support de H_θ , est indépendant de tout x
- (ii) les dérivées partielles suivantes existent et sont finies:

$$\frac{\partial h(\theta)}{\partial \theta_i}, \qquad \frac{\partial^2 h(\theta)}{\partial \theta_i \partial \theta_j}, \qquad \frac{\partial^3 h(\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k}$$

- (iii) la matrice d'information de Fisher

$$I_X^h(\theta) = [E\{\frac{\partial}{\partial \theta_i}\log h(\theta).\frac{\partial}{\theta_j}\log h(\theta)\}]_{i,j=1,\ldots,M}$$

est définie positive.

Considérons un échantillon de taille n et $E_1,E_2,\ldots,E_M$ la partition en M classes corespondante. On peut alors calculer la probabilité observée et associée à chaque classe E_i, en posant:

$$f^* = (f_1^*,f_2^*,\ldots,f_M^*) \quad \text{où} \quad f_i^* = \frac{1}{n}\sum_{j=1}^{n} I_{E_i}(X_j) \quad i = 1,2,\ldots,M \tag{5.5}$$

avec

$$I_{E_i}(X_j) = \begin{cases} 1 & si\ X_j \in E_i \\ 0 & sinon \end{cases}$$

Pour évaluer l'écart entre les fréqences observées et les probabilités théoriques, (le paramètre θ étant supposé inconnu), on propose d'utiliser la mesure d'information $\Delta_r[f,\ h(\theta)]$ définie ci-dessus, avec : $f = \mathbf{E}(f^*)$.

La statistique construite à partir de la mesure $\Delta_r[f,\ h(\theta)]$ est obtenue en remplaçant θ par son estimateur $\widehat{\theta}$. On a donc :

$$\widehat{\Delta}_r = \widehat{\Delta}_r[f,\ h(\theta)] = \Delta_r[f,\ h(\widehat{\theta})] \tag{5.6}$$

Remarque :

Cette mesure de divergence Δ_r définie ci-dessus est dans un sens, une modification de la mesure de divergence de Jeffreys résultant du remplacement de $h(\theta)$ par f ; mais elle semble cependant mieux adapté pour certains types de test, comme nous le verrons plus tard.

On a en effet :

$$\begin{aligned} \lim_{r\to 1} \Delta_r[f,h(\theta)] &= \Delta_1[f,h(\theta)] \\ &= \sum_i f_i \log \frac{f_i}{h_i(\theta)} + \sum_i h_i(\theta) \log \frac{h_i(\theta)}{f_i} \end{aligned} \tag{5.7}$$

On peut maintenant chercher à déterminer la loi de cet estimateur, lorsque l'on considère des échantillons de grandes tailles.

5.2.2 Comportement asymptotique de l'estimateur $\widehat{\Delta}_r[f,\ h(\theta)]$

Nous supposons que l'estimateur $\widehat{\theta}$ de θ vérifie le principe de la normalité asymptotique, à savoir :

$$\sqrt{n}(\widehat{\theta} - \theta) \longrightarrow N[0,\ \Omega] \tag{5.8}$$

où Ω est l'inverse d'une matrice inversible. Si l'on suppose que $\widehat{\theta}$ est obtenue par la méthode du maximum de vraisemblance, on a alors :

$$\Omega(\theta) = I^{-1}(\theta)$$

où $I(\theta)$ représente l'information de Fisher sur le paramètre θ.

Le comportement asymptotique de la statistique $\hat{\Delta}_r[f,\ h(\theta)]$ est donné par le théorème suivant :

Suivant que l'on considère l'hypothèse $f_i = h_i$ ou $f_i \neq h_i$, la loi asymptotique de Δ_r sera donnée par les deux théorèmes suivants :

Théorème 3 :

Soit $\hat{\Delta}_r[f,\ h(\theta)]$ l'estimateur de $\Delta_r[f,\ h(\theta)]$ obtenu en remplaçant θ par l'estimateur $\hat{\theta}$ vérifiant (5.8)

si $f_i = h_i(\theta),\ i = 1,2,\ldots,M$, avec : $f_i = \mathbf{E}(f_i^*)$ on a alors :

$$\frac{n}{r}\hat{\Delta}_r[f,\ h(\theta)] \xrightarrow{L} \chi^2_k$$

où $k = dim\Theta$

Preuve : Pour la démonstration, on pourra se référer à l'annexe 1.

Théorème 4 :

Soit $\hat{\Delta}_r[f,\ h(\theta)]$ l'estimateur de $\Delta_r[f,\ h(\theta)]$ obtenu en remplaçant θ par l'estimateur $\hat{\theta}$ vérifiant (5.8).

si $f_i \neq h_i(\theta),\ i = 1,2,\ldots,M$, avec : $f_i = \mathbf{E}(f_i^*)$ on a alors :

$$\sqrt{n}\{\hat{\Delta}_r[f,\ h(\theta)] - \Delta_r[f,\ h(\theta)]\} \xrightarrow{L} N[0,\Gamma^2]$$

$$\Gamma^2 = \Big(\frac{1}{r-1}\Big)^2 \sum_{i=1}^{k} \Omega^{-1}(\theta_i)\Big\{\sum_{l=1}^{M}\Big((1-r)\frac{f_l^r}{h_l^r(\theta)} + r\frac{h_l^{r-1}(\theta)}{f_l^{r-1}}\Big)\frac{\partial}{\partial\theta_i}h_l(\theta)\Big\}^2$$

Preuve : Voir annexe 2.

5.3 Application aux tests d'adéquation

Nous nous proposons, dans cette section, de mettre en œuvre une procédure de test d'adéquation fondée sur Δ_r, pour ensuite tenter de la comparer par rapport aux tests usuels.

5.3.1 Ajustement à un modèle donné

Pour étudier un phénomène décrit par une variable aléatoire X, on procède à n expériences identiques et indépendantes. On obtient M évènements $E_1,E_2,\ldots,E_M$, chacun d'eux correspondant à une ou plusieurs des valeurs que peut prendre la variable aléatoire X. On veut savoir si la loi empirique constatée de X, peut être assimilée à la loi d'un modèle paramétrique H_θ donné.

Soient $f = (f_1,f_2,\ldots,f_M)$ et $H(\theta) = (h_1(\theta),h_2(\theta),\ldots,h_M(\theta))$ les vecteurs de probabilité définis en (5.4) et (5.5) et correspondant respectivement aux fréquences empiriques et théoriques associées à la partition considérée.

Les hypothèses à tester peuvent être formulées comme suit:

$$\begin{aligned} H_o &: \ f = h \\ H_1 &: \ f \neq h \end{aligned} \tag{5.9}$$

Pour résoudre ce problème de test, on considère la statistique

$$\widehat{\Delta}_r[f,\, h(\theta)] = \frac{1}{r-1}\Big[\sum_{l=1}^{M}\Big\{\frac{f_l^r}{h_l^{r-1}(\widehat{\theta})} + \frac{h_l^r(\widehat{\theta})}{f_l^{r-1}}\Big\} - 2\Big] \tag{5.10}$$

pour estimer l'écart entre la distribution empirique et la loi du modèle.
Ainsi sous l'hypothèse nulle, $\widehat{\Delta}_r$ a tendance à prendre de "petites valeurs", de sorte que, si on se fixe un niveau de signification égal à α, la fonction de test est définie de la manière suivante:

$$\phi_1(x_1,x_2,\ldots,x_n) = \begin{cases} 1 \ si \ \widehat{\Delta}_r > C_\alpha \\ \\ 0 \ \ sinon \end{cases} \tag{5.11}$$

La constante C_α se calcule en utilisant le théorème (3) qui stupile que sous l'hypothèse nulle, la loi de $\frac{n}{r}\Delta_r$ suit asymptotiquement une loi du Khi-deux :

On a alors :

$$C_\alpha = \frac{r}{n}\chi_k^2(\alpha) \tag{5.12}$$

où $\chi_k^2(\alpha)$ est la valeur de la loi du khi-deux pour laquelle, la probabilité d'être dépassée est égale à α.

Sous l'alternative la distribution asymptotiquement normale de $\widehat{\Delta}_r$ est donnée par le théorème (4), ce qui nous permet de formuler l'expression de la puissance comme suit :

$$\begin{aligned} P_n^r &= Prob[\widehat{\Delta}_r > C_\alpha / H_1] \\ &= 1 - \phi[\frac{\sqrt{n}}{\Gamma}(C_\alpha - \Delta_r[f,\, h(\theta)])] \end{aligned}$$

où $\Delta_r[f,\, h(\theta)]$ est la valeur de la divergence entre les fréquences empiriques et théoriques, calculée à partir des observations et $\phi(.)$ désigne la fonction de répartition de la loi normale centrée réduite.

Le test ainsi obtenu est asymptotiquement convergent - au sens de Fraser - puisqu'on a :

$$\lim_{n \to +\infty} P_n^r = 1 \tag{5.13}$$

On obtient ce résultat à partir du fait que C_α tend vers 0 si n tend vers l'infini, et que, dans ce cas, $C_\alpha - \Delta_r$ est négatif, d'où le résultat.

La relation (5.13) signifie que le risque de seconde espèce est asymptotiquement nul.

Une fois le test mis sur pied, nous allons à présent évaluer son degré de performance, en le comparant avec le test du khi-deux et celui de Kolmogorov-Smirnov.

5.3.2 Etude des propriétés des tests par simulation

Pour comparer la précision, où l'exactitude de résultats relatifs à des tests asymptotiques, on dispose en général de deux méthodes fondées respectivement soit sur des procédures d'approximation ou de développement asymptotiques, soit

sur des méthodes d'approximation résultant d'expériences de simulation. Parce que la première méthode conduit souvent à des calculs analytiques compliqués, nous avons choisi de comparer le degré de performance du test fondé sur la statistique $\widehat{\Delta}_r$ ou sur celle du Khi-deux, en procédant à des expériences par le biais des simulations par la méthode de Monte Carlo .

Pour ajuster les données expérimentales, nous allons considérer un processus de génération des données défini par une loi exponentielle $Exp(1/\theta)$, de densité $f(x, \theta) = \frac{1}{\theta} exp(-\frac{x}{\theta}) \mathbf{1}_{\mathbf{R}^+}(x)$.

On pose :

$$\begin{cases} \theta = \theta_o + \epsilon \\ \theta_o > 0, \quad \epsilon \geq 0 \end{cases}$$

On suppose qu'on désire tester l'hypothèse nulle : $\theta_o = 1$, et que $\widehat{\Delta}_r$ est définie par la relation (5.10) précédente.

L'espace des observations est partitionné en trois classes contenant les valeurs caractéristiques de la loi $Exp(1)$, à savoir l'origine et l'espérance mathématique $m = 1$:

$$C_1 = [0,\ 0.1[\ \ ;\ \ C_2 = [0.1,\ 1[\ \ et\ \ C_3 = [1,\ \infty[$$

La statistique de Kolmogorov K_n est basée sur la distribution empirique, donnée par :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{]-\infty,\ x[}(x_i)$$

On pose:

$$K_n(x) = \sup_x |F_n(x) - F_o(x)|$$

où F_n et F_o représentent respectivement la fonction de répartition empirique et la fonction de répartition théorique de l'échantillon.

On choisit ici une taille expérimentale d'échantillon égale à 100 et un nombre de répliques N fixé à 5000. Les résultats de la simulation basée sur ces différentes statistiques sont interprétés à partir d'abord des p-valeurs, puis des valeurs de la puissance, ce pour chacun des tests considérés.

a) Comparaison des p-valeurs

Une comparaison des probabilités de rejet sous l'hypothèse nulle, effectuée de façon traditionnelle, consiste à tabuler les résultats obtenus pour quelques valeurs standards du niveau de signification α (1%, 5% ou 10%).

Niveau de signification nominal α	0.10	0.05	0.01
Probabilité de rejet du χ^2	0.135	0.064	0.011
Probabilité de rejet de Kolmogorov	—	0.002	0.000
Probabilité de rejet de $\widehat{\Delta}_{1/2}$	0.143	0.073	0.016
Probabilité de rejet de $\widehat{\Delta}_1$	0.143	0.073	0.016
Probabilité de rejet de $\widehat{\Delta}_2$	0.148	0.082	0.022
Probabilité de rejet de $\widehat{\Delta}_3$	0.161	0.091	0.030

Tableau 1: *Comparaison entre niveaux de signification nominaux et réponses obtenues dans le cas d'une loi exponentielle.*

Une interprétation en termes de p-valeur, peut être obtenue en utilisant la méthode graphique dont l'essentiel a été rappeler dans le chapitre précédent. Dans ce contexte précis, si $\widehat{F}(x)$ désigne l'estimateur empirique des p-valeurs,

$$\widehat{F}(x) \equiv \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{(p_j \leq x)},$$

$\widehat{F}(x) - x$ traduit la différence entre le niveau de signification estimé par $\widehat{F}(x)$ et le niveau nominal x. On peut donc tracer la courbe correspondante (figure 1) de $(\widehat{F}(x) - x)$ en fonction de x. Pour des raisons liées aux difficultés de calcul des fractiles de la loi de Kolmogorov, nous nous limiterons ici aux statistiques $\Delta_{1/2}$, Δ_1, Δ_2, Δ_3 et celle du khi-deux.

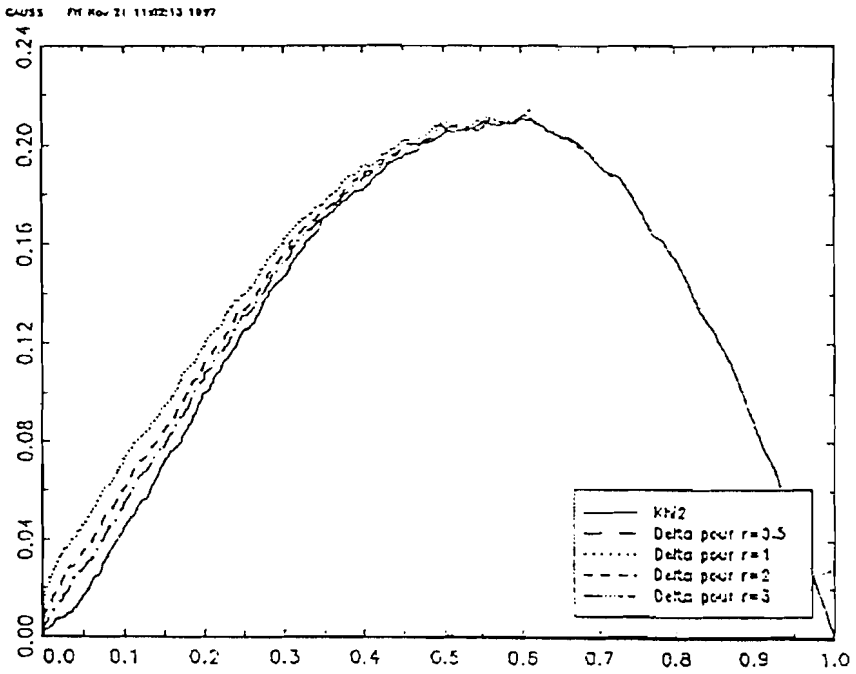

Fig:7 Graphe de $(\widehat{F}(x) - x)$ *en fonction de* x*, pour* $n = 100$

La figure 7 montre, comme l'on pouvait si attendre, compte tenu du tableau 1, que la distance du khi-deux fournit les résultats les plus proches des p-valeurs nominales. On constate, par ailleurs, que l'ensemble des courbes de la figure 7 se comportent globalement de façon analogue (surtout pour $\widehat{\Delta}_{1/2}$ et $\widehat{\Delta}_1$), celle du χ^2 surclassant les autres.

b) Comparaison des puissances

Pour une meilleure appréciation des propriétés de ces statistiques, considérons leur comportement en termes de puissance ; dans le cas de $\widehat{\Delta}_r$, la puissance résultera de l'expression :

$$P_n^r = 1 - \phi[\frac{\sqrt{n}}{\Gamma}(C_\alpha - \Delta_r)]$$

Le tableau qui suit établit la comparaison de ces puissances pour différents tests (basés sur $\widehat{\Delta}_r$, le χ^2 et Kolmogorov).

ϵ	1.1	1.2	1.3	1.4	1.5
khi-deux	0.199	0.358	0.555	0.731	0.855
Kolmogorov	0.014	0.118	0.472	0.905	0.999
$\widehat{\Delta}_{1/2}$	0.130	0.286	0.530	0.769	0.920
$\widehat{\Delta}_1$	0.157	0.331	0.581	0.809	0.940
$\widehat{\Delta}_2$	0.167	0.344	0.594	0.817	0.943
$\widehat{\Delta}_3$	0.177	0.353	0.601	0.821	0.945

Tableau 2: *Valeurs de la puissance en fonction du paramètre ϵ et de la statistique utilisée.*

La puissance du test fondé sur $\widehat{\Delta}_r$ dépend évidemment de r et il apparaît intéressant de se faire une idée de son comportement par rapport à ce paramètre. C'est ce qu'un examen du tableau 2 permet de faire à travers le choix (justifié) d'un ensemble de quatre valeurs de r : $\{0.5; 1; 2; 3\}$. On note une croissance monotone de la puissance avec r. Ainsi, pour un niveau de signification de 5%, la puissance du test du khi-deux, lorsque $\epsilon = 1.4$, est égale à 73.10% alors qu'elle est de 76.90% pour $\widehat{\Delta}_{1/2}$, de 81.70% pour $\widehat{\Delta}_2$ et 82.10% pour $\widehat{\Delta}_3$; la statistique $\widehat{\Delta}_r$ engendrant en fin de compte une puissance supérieure à celle du khi-deux. En comparaison, la fréquence de rejet est de 90.50% pour le test de Kolmogorov.

Rappelons toutefois que pour qu'une métrique donnée puisse être qualifiée de vraie mesure de distance, il faut nécessairement qu'elle soit positive, symétrique et vérifie l'inégalité triangulaire.

Or parmi les mesures de proximité qui ont été proposées et qui ont conduit à toute une panoplie de mesures divergence, rares sont celles qui sont à proprement parlé des mesures de distance, puisqu 'elles ne vérifient pas en général, l'inégalité triangulaire. On les appelle alors des pseudo-distances. Cependant, dans le cas de la mesure définie à partir de (5.3), lorsque le paramètre $r = 1/2$, on obtient une

vraie distance en posant :

$$\begin{aligned}\Delta_{1/2}[f,h(\theta)] &= D_{1/2}^{1/2}[f,\, h(\theta)] + D_{1/2}^{1/2}[h(\theta),\, f] \\ &= 2\Big\{\sum_{i=1}^{n}(f_i^{1/2} - h_i^{1/2}(\theta))^2\Big\}\end{aligned}$$

Cette mesure est très appréciée en analyse statistique; et cet intérêt se justifie probablement par le fait qu'elle présente la propriété d'être bornée. En effet elle peut également s'écrire sous la forme suivante :

$$\Delta_{1/2}[f,h(\theta)] = 4[1 - \sum_{i=1}^{n}\Big\{f_i\ h_i(\theta)\Big\}^{1/2}] \tag{5.14}$$

Soit alors :

$$0 \leq \Delta_{1/2}[f,h(\theta)] \leq 4$$

Notons que d'autres distances proportionnelles à $\widehat{\Delta}_{1/2}$ ont été définies entre lois de probabilité. On peut citer par exemple, pour deux distributions de densités p et q données, les mesures suivantes :

$$B[p,\, q] = [1 - \sum_{i=1}^{n}\Big\{p_i\ q_i\Big\}^{1/2}] \tag{5.15}$$

$$M[p,\, q] = \sum_{i=1}^{n}\Big\{p_i - q_i\Big\}^2 \tag{5.16}$$

Les expression (5.15) et (5.16) représentent respectivement les distances de Bhattacharya (1943) et de Matusita (1951, 1967), qui vérifient toutes les deux les propriétés d'une vraie métrique.

Ainsi, les propriétés asymptotiques obtenues à partir d'un test fondé sur la distance $\Delta_{1/2}[p,\, q]$ sont pratiquement les mêmes que si on utilise la distance $B[p,\, q]$ ou $M[p,\, q]$. Cela résulte du fait que ces deux dernières mesures constituent des formes déterministes de la divergence $\Delta_{1/2}[p,\, q]$. On obtient en effet les égalités suivantes :

$$\Delta_{1/2}[p,\, q] = 4B[p,\, q] = 2M[p,\, q] \tag{5.17}$$

La discussion ci-dessus, qui souligne la place privilégié de $\Delta_{1/2}$ parmi les mesures de divergences d'une part et la bonne performance aussi bien en liaison avec les p-valeurs mais également par rapport à la puissance d'autre part, nous incite à

proposer, dans la section suivante, une application pour la résolution d'un test de sélection entre modèles paramétriques.

5.4 Test de sélection de modèles

La recherche d'un test pour choisir une distribution parmi deux distributions s'appuie traditionnellement sur la méthode de Akaike (1973) ou celle souvent mieux adaptée de Vuong et Wang (1993) dont la base, dans le dernier cas, est la distance du khi-deux.

Par comparaison, on suggère ici une procédure pour déterminer, entre deux modèles paramétriques H_θ et G_π, celui qui s'adapte le mieux à la loi empirique d'une série d'observations donnée. On se basera pour cela sur les mesures d'information de type $\Delta_{1/2}$ servant de mesure de divergence entre le modèle H_θ ou G_π et les observations.

Les fonctions f, h et g désignent respectivement la fréquence empirique, la loi théorique du modèle H_θ et celle de G_π. Les estimateurs de θ et de π vérifient la relation (5.8).

Soient $\widehat{\Delta}_{1/2}[f,\, h(\theta)]$ et $\widehat{\Delta}_{1/2}[f,\, g(\pi)]$ les estimateurs respectifs de $\Delta_{1/2}[f,\, h(\theta)]$ et $\Delta_{1/2}[f,\, g(\pi)]$.

On considère les hypothèses suivantes :

(i) H_o : $\Delta_{1/2}[f,\, h(\theta)] = \Delta_{1/2}[f,\, g(\pi)]$

(ii) $\mathrm{H}_{1,g}$: $\Delta_{1/2}[f,\, h(\theta)] > \Delta_{1/2}[f,\, g(\pi)]$

(iii) $\mathrm{H}_{1,h}$: $\Delta_{1/2}[f,\, h(\theta)] < \Delta_{1/2}[f,\, g(\pi)]$.

L'hypothèse (i) signifie que les modèles H_θ et G_π sont équivalents ; (ii) traduit le fait que G_π est meilleur que H_θ et (iii) suggère de choisir H_θ plutôt que G_π.

La résolution de ce problème de choix entre H_θ et G_π sera fondée sur la statistique

$$\widehat{D_n} = D_n[h(\widehat{\theta}),\, g(\widehat{\pi})] = \sqrt{n}\{\widehat{\Delta}_{1/2}[f,\, h(\theta)] - \widehat{\Delta}_{1/2}[f,\, g(\pi)]\}$$

qui estime $\sqrt{n}\{\Delta_{1/2}[f,\, h(\theta)] - \Delta_{1/2}[f,\, g(\pi)]\}$.

Sous l'hypothèse nulle H_o, la loi asymptotique de $\widehat{D}_n$ est donnée par une version du théorème de Vuong-Wang (1993) :

Théorème 5 : Si $\hat{\theta}$ et $\hat{\pi}$ représentent respectivement les $E.M.V$ de θ et π , on a (avec la notation ci-dessus de $\widehat{D}_n$) :

(1) sous H_o : $\widehat{D}_n[h(\theta),\, g(\pi)] \longrightarrow N[0,\, \Sigma^2]$

(2) sous $H_{1,g}$: $\widehat{D}_n[h(\theta),\, g(\pi)] \xrightarrow{P} +\infty$

(3) sous $H_{1,h}$: $\widehat{D}_n[h(\theta),\, g(\pi)] \xrightarrow{P} -\infty$

avec :

$$\frac{\partial}{\partial\theta}\Delta_{1/2}(\theta) = p(\theta) \qquad et \qquad \frac{\partial}{\partial\pi}\Delta_{1/2}(\pi) = q(\pi)$$

$$C^t = C^t(\theta,\, \pi) = (p^t(\theta)\ ,\ -q^t(\pi)) \qquad ; \qquad \hat{\eta} = \sqrt{n}(\hat{\theta} - \theta)$$

$$et \qquad \hat{\delta} = \sqrt{n}(\hat{\pi} - \pi)$$

$$\Sigma^2 = C^t \Lambda C$$

où

$$\Lambda = \begin{bmatrix} I^{-1}(\theta) & \mathbf{E}(\hat{\delta}\ \hat{\eta}^t) \\ \mathbf{E}(\hat{\eta}\ \hat{\delta}^t) & I^{-1}(\pi) \end{bmatrix}$$

puisque $\hat{\eta}$ et $\hat{\delta}$ sont des variables aléatoires centrées.

Preuve : Pour la démonstration de ce théorème, on pourra se référer à l'annexe 3.

5.4.1 Règle de décision associée à la statistique $\widehat{D}_n$

Nous allons nous appuyer sur l'inégalité triangulaire que vérifie la métrique $\Delta_{1/2}$ pour un encadrement de $\widehat{D}_n$.

En effet :

$$\Delta_{1/2}[f,\, h(\theta)] \leq \Delta_{1/2}[f,\, g(\pi)] + \Delta_{1/2}[g(\pi),\, h(\theta)]$$

soit :

$$\Delta_{1/2}[f,\, h(\theta)] - \Delta_{1/2}[f,\, g(\pi)] \leq \Delta_{1/2}[g(\pi),\, h(\theta)] \qquad (5.18)$$

D'autre part :

$$\Delta_{1/2}[f,\, g(\pi)] \leq \Delta_{1/2}[f,\, h(\theta)] + \Delta_{1/2}[h(\theta),\, g(\pi)]$$

ce qui entraine :

$$\Delta_{1/2}[f,\, g(\pi)] - \Delta_{1/2}[f,\, h(\theta)] \leq \Delta_{1/2}[h(\theta),\, g(\pi)] \qquad (5.19)$$

Posons :

$$k_n = \sqrt{n}\ \Delta_{1/2}[h(\widehat{\theta}),\, g(\widehat{\pi})]$$

En multipliant (5.18) par $\sqrt{n}$ et (5.19) par $-\sqrt{n}$, et en remplaçant ensuite θ et π par leurs estimateurs respectifs $\widehat{\theta}$ et $\widehat{\pi}$, on obtient en fait :

$$-k_n \leq \widehat{D}_n \leq k_n$$

Pour réaliser le test de choix entre h et g, on peut envisager une règle de décision définie comme suit, pour un niveau de signification supposé égal α :

- il y a équivalence entre h et g si :

$$\widehat{D}_n \in \left[- z_{\alpha/2}\ \Sigma\ ,\ z_{\alpha/2}\ \Sigma \right]$$

- on décide en faveur de h lorsque :

$$\widehat{D}_n \in \left[- k_n\ ,\ - z_{\alpha/2}\ \Sigma \right]$$

- on décide en faveur de g si :

$$\widehat{D}_n \in \left[z_{\alpha/2}\ \Sigma,\ \ k_n \right]$$

Σ^2 représentant la variance de la statistique $\widehat{D}_n$ et $z_{\alpha/2}$ le quantile $(1 - \alpha/2)$ de la loi normale centrée réduite.

5.4.2 Exemples d'application

On propose ici une comparaison entre la statistique $\widehat{D}_n[h(\theta),\, g(\pi)]$, construite avec $\hat{\Delta}_{1/2}$ et $\widehat{K}_n[h(\theta),\, g(\pi)]$, obtenue en fonction de la statistique de Pearson. Ces statistiques sont définies comme suit :

$$\widehat{D}_n[h(\theta),\, g(\pi)] = \sqrt{n}\{\hat{\Delta}_{1/2}[f,\, h(\theta)] - \hat{\Delta}_{1/2}[f,\, g(\pi)]\} \qquad (5.20)$$

$$\widehat{K}_n[h(\theta),\, g(\pi)] = \frac{1}{\sqrt{n}}\{\hat{Q}_n[f,\, h(\theta)] - \hat{Q}_n[f,\, g(\pi)]\} \qquad (5.21)$$

$\hat{Q}_n[f,\, h(\theta)]$ désignant la distance du khi-deux entre la fréquence empirique f et la distribution théorique h.

A titre d'illustration, des simulations par Monte Carlo ont été mises en œuvre à partir de quelques distributions, afin de comparer la méthode de Vuong et Wang avec la procédure que nous avons proposée. On se limitera ici, à trois types de lois dont les densités de probabilité sont définies sur :

- un intervalle $[a,\, b]$,
- l'ensemble R_+
- l'ensemble R.

Le nombre de répliques utilisé pour construire les distributions empiriques est $N = 5000$ et la taille des échantillons considérés varie entre 70 et 800. Le niveau de signification retenu est de 5%.

Cas de deux distributions définies sur $[0,\, 1]$

On veut sélectionner un modèle parmi deux distributions (une loi Bêta et une loi uniforme), sur la base d'une série d'observations obtenues à partir de deux processus de génération des données (PGD) Y_1 et Y_2 de densités respectives f_1 et f_2 :

$$Y_1 \sim Be(p,\, q) \qquad f_1 = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1}\mathbf{1}_{[0,\,1]}(x)$$

$$Y_2 \sim U_{[0,\,1]} \qquad f_2 = \mathbf{1}_{[0,\,1]}(x)$$

Afin d'espérer des résultats tangibles, il est nécessaire de pouvoir raisonnablement discerner les deux distributions ; on choisira à cet effet les valeurs $p = 1$ et $q = 2$ et l'on regroupera les données en trois classes $C_1 = [0, 0.2[$; $C_2 = [0.2, 0.8[$ et $C_3 = [0.8, 1]$.

On obtient les tableaux suivants :

PGD : $Y_1 \sim Be(1, 2)$

Taille de l'échantillon		70	100	150	200
modèle fondé sur	décision: f_1	0.634	0.809	0.940	0.981
la statistique $\widehat{K}_n$	indécision	0.366	0.191	0.060	0.019
	décision: f_2	0.000	0.000	0.000	0.000
Modèle fondé sur	décision: f_1	0.605	0.806	0.946	0.987
la statistique $\widehat{D}_n$	indécision	0.395	0.194	0.054	0.013
	décision: f_2	0.000	0.000	0.000	0.000

Tableau 3 : Comparaison des probabilités d'acceptation pour $\widehat{K}_n$ et $\widehat{D}_n$

PGD : $Y_2 \sim U_{[0,\ 1]}$

Taille de l'échantillon		100	150	200	300
modèle fondé sur	décision: f_1	0.000	0.000	0.000	0.000
la statistique $\widehat{K}_n$	indécision	0.449	0.254	0.123	0.024
	décision: f_2	0.551	0.746	0.877	0.976
Modèle fondé sur	décision: f_1	0.000	0.000	0.000	0.000
la statistique $\widehat{D}_n$	indécision	0.164	0.050	0.014	0.001
	décision: f_2	0.836	0.950	0.986	0.999

Tableau 4 : Comparaison des probabilités d'acceptation pour $\widehat{K}_n$ et $\widehat{D}_n$

Cas de deux distributions définies sur R_+

Envisageons le problème qui consiste à choisir, par exemple, entre une distribution exponentielle $Exp(\theta)$ de paramètre θ et une loi Gamma $\Gamma(p, \alpha)$ de paramètres

(p, α), de densités respectives :

$$\begin{cases} f_3(x, \theta) = \theta exp(-\theta x) \\ x \geq 0 \quad et \quad \theta > 0 \end{cases} \tag{5.22}$$

$$\begin{cases} f_4(x, p, \alpha) = \dfrac{\alpha^p}{\Gamma(p)} x^{p-1} exp(-\alpha x) \\ x \geq 0, \quad p > 0 \quad et \quad \alpha > 0 \end{cases} \tag{5.23}$$

Nous supposerons dans ce qui suit que les estimateurs de θ et α sont obtenus par la méthode MV.

Pour des raisons de calcul, on donnera une valeur entière à p, la valeur 2 par exemple. Par ailleurs, pour espérer ici obtenir des résultats significatifs, on prendra dans (5.22) $\theta = 0.7$ et dans (5.23) $\alpha = 1$, de telle sorte que les données issues de ces deux lois conduisent à la même variance, atténuant ainsi " l'écart " entre les distributions choisies. Les observations seront réparties en trois classes :

$C_1 = [0, 0.3[$; $C_2 = [0.3, 0.1.5[$; $C_3 = [1.5, +\infty[$

Dans le cas présent, nous générons les échantillons à partir de deux processus de génération des données :

$Y_3 \sim Exp(0.7)$

$Y_4 \sim \Gamma[2, 1]$

PGD : $Y_3 \sim Exp(0.707)$

Taille de l'échantillon		100	200	300	500
modèle fondé sur	décision: f_3	0.449	0.749	0.906	0.989
la statistique $\widehat{K}_n$	indécision	0.551	0.251	0.094	0.011
	décision: f_4	0.000	0.000	0.000	0.000
Modèle fondé sur	décision: f_3	0.578	0.859	0.961	0.997
la statistique $\widehat{D}_n$	indécision	0.422	0.141	0.039	0.003
	décision: f_4	0.000	0.000	0.000	0.000

Tableau 5 : Comparaison des probabilités d'acceptation pour $\widehat{K}_n$ et $\widehat{D}_n$

PGD : $Y_4 \sim \Gamma[2, 1]$

Taille de l'échantillon		300	500	600	800
modèle fondé sur	décision: f_3	0.003	0.001	0.000	0.000
la statistique $\widehat{K}_n$	indécision	0.681	0.464	0.387	0.249
	décision: f_4	0.316	0.535	0.613	750
Modèle fondé sur	décision: f_3	0.000	0.000	0.000	0.000
la statistique $\widehat{D}_n$	indécision	0.502	0.271	0.207	0.092
	décision: f_4	0.497	0.729	0.793	0.908

Tableau 6 : Comparaison des probabilités d'acceptation pour $\widehat{K}_n$ et $\widehat{D}_n$

Cas de deux distributions définies sur R

On veut choisir entre une distribution de Laplace $\xi(\alpha, \lambda)$ (ou loi exponentielle double) et une loi normale $N[m, \sigma^2]$. On considère les PGD Y_5 et Y_6 suivants :

$$Y_5 \sim \xi(\alpha, \lambda) \qquad f_5(x, \alpha, \lambda) = \frac{\lambda}{2} e^{-\lambda|x-\alpha|}$$

$$Y_6 \sim N[m, \sigma^2] \qquad f_6(x, m, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Nous supposerons les observations issues de populations ayant même moyenne, $m = \alpha = 0$ et, pour simplifier, on prendra $\sigma = \lambda = 1$. On pose $C_1 =]-\infty, -2[$; $C_2 = [-2, 2[$ et $C_3 = [2, -\infty[$, comme partition associée aux observations.
On obtient les résultats ci-dessous :

PGD : $Y_5 \sim \xi(1, 1)$

Taille de l'échantillon		100	200	300	500
modèle fondé sur	décision: f_5	0.279	0.467	0.652	0.865
la statistique $\widehat{K}_n$	indécision	0.721	0.533	0.348	0.135
	décision: f_6	0.000	0.000	0.000	0.000
Modèle fondé sur	décision: f_5	0.311	0.599	0.797	0.948
la statistique $\widehat{D}_n$	indécision	0.687	0.401	0.203	0.052
	décision: f_6	0.002	0.000	0.000	0.000

Tableau 7 : Comparaison des probabilités d'acceptation pour $\widehat{K}_n$ et $\widehat{D}_n$

PGD : $Y_6 \sim N[0, 1]$

Taille de l'échantillon		70	100	150	400
modèle fondé sur	décision: f_5	0.017	0.010	0.003	0.000
la statistique $\widehat{K}_n$	indécision	0.818	0.702	0.398	0.036
	décision: f_6	0.165	0.288	0.599	0.964
Modèle fondé sur	décision: f_5	0.000	0.000	0.000	0.000
la statistique $\widehat{D}_n$	indécision	0.728	0.634	0.507	0.088
	décision: f_6	0.272	0.366	0.493	0.912

Tableau 8 : Comparaison des probabilités d'acceptation pour $\widehat{K}_n$ et $\widehat{D}_n$

Dans le tableau 3, le test fondé sur la statistique du khi-deux donne des résultats sensiblement proches de celui fondé sur la mesure D_n. Dans les tableaux 4, 5, 6 et 7, les résultats sont nettement meilleurs lorsque l'on considère le test obtenu à partir de $\widehat{D}_n$.

En revanche, dans le tableau 8, on notera que la méthode associée à $\widehat{K}_n$ semble préférable dès que la taille de l'échantillon devient suffisamment grande . En effet, sur la base de 400 observations par exemple, la bonne décision se traduit par une probabilité d'acceptation de l'ordre de 96.40% pour $\widehat{K}_n$ et de 91.20% pour $\widehat{D}_n$.

Conclusion

Nous avons, dans ce chapitre, tenté d'utiliser une distance informationnelle de type Rényi, pour des tests aussi bien d'ajustement que de choix de modèles. Pour en cerner l' efficacité, nous avons en parallèle, comparé nos résultats dans les deux situations avec ceux fournis par les tests classiques du khi-deux ou de Kolmogorov. De cette tentative informationnelle et de cette comparaison, on retiendra essentiellement ce qui suit :

- pour le test d'ajustement, à travers le critère des p-valeurs, les distances $\Delta_{1/2}$ et Δ_1 (très proches l'une de l'autre) sont, parmi les Δ_r, les plus efficaces mais s'avèrent moins performantes que le khi-deux (le test de Kolmogorov n'a pas été ici pris en compte en raison de difficultés de calcul évidentes). Avec le critère "puissance", à partir de certaines valeurs du paramètre ϵ, Δ_r quel que soit r est préférable au

khi-deux, le test de Kolmogorov s'avérant cependant meilleur ;

• pour le test de choix de modèle, on s'est limité, en le justifiant, à comparer $\widehat{\Delta}_{1/2}$ et le khi-deux au travers des statistiques $\widehat{D}_n$ et $\widehat{K}_n$ données en (5.20) et (5.21). Il apparaît, d'après les résultats obtenus, qu' aucune des deux statistiques de test considérées ici n'est systématiquement plus performante que l'autre (tableaux 3 et 8). Cependant, dans de nombreux cas, le test basé sur $\widehat{\Delta}_{1/2}$ engendre une meilleure puissance (tableaux 4, 5, 6 et 7).

On retiendra enfin que dans le cadre des petits échantillons (pour les échantillons de grande taille, ces statistiques de test sont équivalentes), les résultats obtenus, en plus de la simplicité de calcul de $\widehat{\Delta}_{1/2}$, plaident en faveur de cette distance dans plusieurs situations.

$\triangle$

Troisième partie

Interprétation de la distribution généralisée

Chapitre 6

Approche bayesienne fondée sur la distribution généralisée

6.1 Introduction

Dans l'analyse bayésienne, un modèle est composé d'un espace d'échantillonnage X, un espace des paramètres Θ et une distribution conjointe définie sur ces deux espaces. Supposons donc, en sus de la loi des observations, $f(x/\theta)$, une distribution a priori π définie sur θ.
On peut alors écrire, à partir de ces distributions

(1) la loi conjointe de (θ, x),

$$\psi(x, \theta) = f(x/\theta)\pi(\theta) \ ;$$

(2) la loi marginale de x,

$$\begin{aligned} m(x) &= \int \psi(x,\theta)d\theta \\ &= \int f(x/\theta)\pi(\theta)d\theta \end{aligned}$$

(3) la loi a postériori de θ, obtenu par la formule de Bayes,

$$\pi(\theta/x) = \frac{f(x/\theta)\pi(\theta)}{\int f(x/\theta)\pi(\theta)d\theta}$$

$$= \frac{f(x/\theta)\pi(\theta)}{m(x)} \qquad (6.1)$$

C'est en se fondant sur la notion de distribution généralisée suggérée dans la section (2.5.2), que nous proposons une approche bayésienne de la distribution d'ordre α.

Dans le cadre présent, partant d'une loi de densité $f(x/\theta)$, on s'intéresse à $f^{\alpha}(x/\theta)$, renormalisée à l'unité au travers de $\phi_{\alpha}(\theta/x)$:

$$\phi_{\alpha}(\theta/x) = \frac{f^{\alpha}(x/\theta)}{\int_{\Theta} f^{\alpha}(x/\theta)d\theta}$$

L'interprétation de l'α-distribution en termes bayésiens, repose sur le choix de la loi a priori. Nous considérons ici, l'approche de l'inférence bayésienne qui consiste à privilégier, sous coût quadratique, le recours à la loi a postériori, qui représente l'actualisation de l'information a priori, $\pi(\theta)$, au vu de l'information contenue dans les observations à partir de $f^{\alpha}(x/\theta)$.

On pose :

$$\pi_{\alpha}(\theta/\ x) = \frac{f^{\alpha}(x/\theta)\ \pi(\theta)}{\int_{\Theta} f^{\alpha}(x/\theta)\ \pi(\theta)d\theta}$$

Lorsque l'on est en présence d'une situation où on ne dispose d'aucune information sur le modèle, il est possible de bâtir une distribution a priori qui intégre notre ignorance sur le paramètre du modèle. De telles méthodes, connues sous le nom de loi a priori non informative, conduisent, notamment lorsque le paramètre appartient à $]-\infty,\ +\infty[$, à retenir une loi a priori contante $\pi(\theta) = c$ sur un intervalle $[a, b]$. On pourra considérer que la vraisemblance a priori d'un intervalle $[a, b]$ est proportionnelle à sa longueur $b - a$, donc la loi a priori, dans ce cas, est précisément la mesure de Lebesgue sur R.

Dans notre contexte, cela se traduit par :

$$\begin{aligned} \pi_{\alpha}(\theta/\ x) &= \phi_{\alpha}(\theta/\ x) \\ &= \frac{f^{\alpha}(x/\theta)}{\int_{\Theta} f^{\alpha}(x/\theta)d\theta} \end{aligned} \qquad (6.2)$$

Lorsque α prend la valeur 1, les estimateurs obtenus à partir de (6.2) donnent

de bonnes performances et correspondent en général à des estimateurs classiques comme ceux issus de la méthode du M. V, ce qui justifie cette extension.

Exemple :

Si $X \sim N[\theta,\sigma^2]$ et $\pi(\theta) = c$, alors $\pi_\alpha(\theta/x) \sim N[\theta,\frac{\sigma^2}{\sqrt{\alpha}}]$.

Ainsi la moyenne a posteriori redonne l'estimateur du maximum de vraisemblance X.

La généralisation (6.2) n'est pas génante, même si l'interprétation de α reste primordiale, puisque, suivant le principe de vraisemblance, seule la loi a postériori est importante. Cette approche bayésienne suggère en effet, de travailler conditionnellement aux observations, ce qui se traduit par une inversion des probabilités par rapport à l'approche fréquentiste, tout en restant fidèle au principe de Vraisemblance que nous rappelons ci-dessous.

Principe de vraisemblance

Toute l'inférence sur θ tirée de x est contenue dans la vraisemblance $\ell(\theta/x)$. De plus si x_1 et x_2 sont tels qu'il existe une constante C telle que, pour tout θ,

$$\ell(\theta/x_1) = C\ell(\theta/x_2),$$

ils apportent la même information sur θ et doivent conduire à la même inférence.

Notons au passage que le principe de vraisemblance n'est valide que si d'une part l'inférence concerne le même paramètre θ, et d'autre part θ prend en compte toutes les inconnues du modèle.

6.2 Résolution numérique des équations de Vraisemblance à partir de ϕ_α

Plaçons nous dans un cadre purement statistique. Considérons le modèle qui met en jeu trois espaces : Ξ espace des observations, Θ, espace des paramètres, et

$\mathcal{D}$, espace des décisions. L'inférence consiste donc à prendre une décision δ concernant $\theta \in \Theta$ au vu d'une série d'observations $y = (y_1, \ldots, y_n)$, y et θ étant reliés par la densité $f(y/\theta)$.

L'estimation par la méthode du Maximum de Vraisemblance, consiste, une fois l'observation $y = (y_1, \ldots, y_n)$ effectuée, à retenir, comme estimation du paramètre θ, une valeur $\delta = \widehat{\theta}(y)$ rendant maximale la fonction de vraisemblance :

$$\theta \longrightarrow l(y/\theta) = \prod_{i=1}^{n} f(y_i/\theta)$$

Ainsi, par définition, on appelle estimateur du maximum de vraisemblance du paramètre θ, la solution du problème de maximisation :

$$\max_{\theta \in \Theta} l(Y/\theta)$$

Remarquons au passage que la définition précédente pose divers problèmes, liés au fait que :

- (a) la solution du problème de maximisation peut ne pas exister,
- (b) il peut exister une multiplicité de solutions.
- (c) la fonction de vraisemblance peut ne pas être définie de manière unique.

Par ailleurs, la démarche à suivre pour une mise en œuvre pratique des estimateurs du maximum de vraisemblance, conduit généralement à deux situations :

1- une résolution analytique, qui consiste à rechercher les solutions des équations de vraisemblance :

$$\nabla l(y_1, \ldots, y_n/\theta) = 0$$

2- une résolution numérique, qui représente une alternative à (1), lorsque précisément, une résolution analytique devient impossible. Cette approche est basée sur divers algorithmes comme les méthodes de Gradients, la procédure de Newton-Raphson, ou encore l'algorithme espérance-maximisation entre autres.

Dans le cadre de l'approche par approximation, il est possible de proposer une réponse bayésienne pour une résolution numérique des équations de vraisemblance. L'idée de base repose principalement sur le lemme suivant :

Lemme :

Si $\pi(\theta)$ est une densité de probabilité sur Θ telle que :

$$supp(\pi) \supset supp\ l(y_1 \ldots, y_n/\theta)$$

on a alors

$$\pi_k(\theta/y_1, \ldots, y_n) \propto \pi(\theta) l^k(y_1, \ldots, y_n/\theta)$$

et π_k tend vers une masse de Dirac en $\widehat{\theta}$ quand k tend vers $+\infty$

D'autre part, on peut noter que la plupart des lois usuelles appartiennent à une famille de lois de probabilités dont l'importance - surtout liée aux statistiques exhaustives - est essentielle notamment du fait de l'existence de résultats généraux en théorie de l'estimation et en théorie des tests. De telles lois sont dites *" familles exponentielles "* et sont étudiées en détail dans Lehmann (1983 et 1986).

Définition :

Soient μ, mesure $\sigma-$finie sur Ξ, et Θ l'espace des paramètres. On définit h et C, respectivement fonctions de Ξ et Θ dans $\mathbf{R}_+$, et R et T fonctions de Θ et Ξ dans R^k. La famille de distributions de densité (par rapport à μ)

$$f(x/\theta) = C(\theta) H(x) \exp\{R(\theta) T(x)\} \qquad (6.3)$$

est dite *"famille exponentielle de dimension k"*. Dans le cas particulier où $\Theta \subset R^k$, $\Xi \subset R^k$ et

$$f(x/\theta) = C(\theta) h(x) \exp(\theta x), \qquad (6.4)$$

la famille est dite *naturelle*.

L'expression (6.3) peut également s'écrire sous la forme :

$$\log f(x, \theta) = \sum_{j=1}^{k} r_j(\theta) T_j(x) + b(x) + \beta(\theta) \qquad (6.5)$$

Dans le cadre de cette famille de lois, les distributions d'ordre α associées à une famille *"mère"* tiennent une place privilégiée. Ainsi, lorsque la densité de probabilité f est supposée appartenir à l'ensemble des puissances α^{ieme} intégrable, on introduit le terme :

$$K_\alpha(x) = \int_\Theta f^\alpha(x, \theta) d\theta$$

correspondant au facteur normatif de la distribution généralisée.

Cette distribution d'ordre ϕ_α, associée à $f(x, \theta)$ prend alors la forme simple suivante :

$$\log \phi_\alpha = A(x)R(\theta) + B(x) + C(\theta) \qquad (6.6)$$

avec:

$$A(x) = \alpha T(x); \quad B(x) = \alpha b(x) - \log K_\alpha(x); \quad C(\theta) = \alpha\beta(\theta)$$

La famille exponentielle est donc stable par extension à l'ordre α, en ce sens qu'on obtient une forme canonique des lois à résumé exhaustif pour cette opération.

L'expression de la distribution généralisée apparaît donc, dans ce contexte, comme une réactualisation de la loi mère associée, au travers d'une reparamétrisation du modèle.

Dans ce contexte, on peut interpréter la densité

$$\phi_\alpha(\theta/y) = \frac{f^\alpha(y/\theta)}{\int f^\alpha(y/\theta) d\theta}$$

comme loi a postériori de θ.

On remarquera que même lorsque la loi a priori $\pi(\theta)$ n'est pas constante, on peut toujours la choisir en posant :

$$\pi_\alpha(\theta) = \frac{f^{\alpha-1}(y/\theta)}{\int_\Theta f^{\alpha-1}(y/\theta) d\theta}$$

Ce qui permet par conséquent, d'établir la relation suivante :

$$\phi_\alpha(\theta/y) \propto f(y/\theta)\pi_\alpha(\theta)$$

L'établissement d'un algorithme numérique fondé sur une analyse bayésienne, permet, en s'appuyer sur le lemme précédent, de résoudre les équations de vraisemblance, en considérant l'espérance mathématique de la loi a postériori ϕ_α.

Ainsi, d'après le lemme précédent, on a :

$$\begin{aligned} \mathbf{E}^{\phi_\alpha}(\theta) &= \frac{\int_\Theta \theta f^\alpha(y/\theta)d\theta}{\int_\Theta f^\alpha(y/\theta)d\theta} \\ &= \int_\Theta \theta\phi_\alpha(\theta/y)d\theta \overset{\alpha\to+\infty}{\longrightarrow} \widehat{\theta} \end{aligned}$$

Exemple 1 : (Régression logistique)

Soit $(y_1,x_1)\ldots,(y_n,x_n)$ ou

$$y_i = \begin{cases} 0 & \text{avec une probabilité} \quad 1-\rho_i \\ 1 & \text{avec une probabilité} \quad \rho_i \end{cases}$$

avec

$$\rho_i = \frac{e^{\lambda^t x_i}}{1+e^{\lambda^t x_i}} \qquad \lambda \in R^k \quad ; \quad x_i \in R^k$$

La densité conditionnelle de y_i sachant x_i est donnée par :

$$\begin{aligned} f(y_i/x_i,\lambda) &= \frac{e^{\lambda^t(\sum_{i=1}^n x_i y_i)}}{\prod_i^n (1+e^{\lambda^t x_i})} \\ &= e^{\lambda^t(\sum_{i=1}^n x_i y_i) - \psi(\lambda)} \end{aligned} \tag{6.7}$$

où

$$\psi(\lambda) = \sum_{i=1}^n \log(1+e^{\lambda^t x_i})$$

La solution $\widehat{\lambda}$ vérifie la relation suivante :

$$\sum_{i=1}^n \frac{x_i e^{\widehat{\lambda}^t x_i}}{1+e^{\widehat{\lambda}^t x_i}} = \sum_{i=1}^n x_i y_i$$

L'équation de vraisemblance associée à ce modèle est non linéaire par rapport au paramètre λ ; il n'est pas possible d'exprimer l'estimateur comme fonction

simple des observations et l'équation devra donc être résolue au moyen d'algorithme numérique.

Une réponse bayésienne, fondée sur la distribution généralisée

$$\phi_\alpha(\lambda/x,y) \propto e^{\alpha\lambda^t(\sum_{i=1}^{n} x_i y_i) - \alpha\psi(\lambda)}$$

revient à considérer la moyenne a postériori du paramètre λ :

$$\mathbf{E}^{\phi_\alpha}(\lambda) = \frac{\int \lambda e^{\alpha\lambda^t(\sum_{i=1}^{n} x_i y_i) - \alpha\psi(\lambda)} d\lambda}{\int e^{\alpha\lambda^t(\sum_{i=1}^{n} x_i y_i) - \alpha\psi(\lambda)} d\lambda}$$

On doit donc effectuer deux intégrations numériques pour calculer numérateur et dénominateur. On pourra alors utiliser la méthode de Simpson (au travers du polynôme de Hermite de degré n) ou, lorsque la dimension de Θ augmente, il est préférable de se référer à une méthode de simulation.
La méthode Monte-Carlo, de fonction d'importance η (où η est une densité de probabilité arbitraire sur λ), propose une solution par approximation de l'intégrale

$$\int \lambda e^{\lambda^t(\sum_{i=1}^{n} x_i y_i) - \psi(\lambda)} d\lambda \qquad (6.8)$$

L'algorithme consiste ici, à générer $\lambda_1, \ldots, \lambda_n$ suivant η, et (6.8) est alors approchée par :

$$\frac{1}{N}\sum_{i=1}^{N} \lambda_i \Omega(\lambda_i)$$

avec :

$$\Omega(\lambda) = \frac{exp\ \alpha\lambda^t(\sum_{i=1}^{n} x_i y_i) - \alpha\psi(\lambda)}{\eta(\lambda)}$$

La loi des grands nombres assure, sous certaines conditions de régularité, que cette approximation converge vers (6.8) quand N tend vers l'infini :

$$\frac{1}{N}\sum_{i=1}^{N} \lambda_i \Omega(\lambda_i) \longrightarrow \int \lambda \Omega(\lambda)\eta(\lambda) d\lambda$$

On obtient :

$$\mathbf{E}^{\phi_\alpha}(\lambda) = \frac{\int \lambda e^{\alpha\lambda^t(\sum_{i=1}^{n} x_i y_i) - \alpha\psi(\lambda)} d\lambda}{\int e^{\alpha\lambda^t(\sum_{i=1}^{n} x_i y_i) - \alpha\psi(\lambda)} d\lambda} \approx \frac{\sum_{i=1}^{n} \lambda_i \Omega(\lambda_i)}{\sum_{i=1}^{n} \Omega(\lambda_i)} \qquad (6.9)$$

En général, le résultat (6.9) peut être obtenu de manière relativement rapide. Mais lorsque le calcul des intégrales associées est complexe, on peut toujours recourir à des méthodes de simulation.

Il suffit alors de déterminer la limite de $E^{\phi_\alpha}(\lambda)$ lorsque α tend vers l'infini, pour obtenir un estimateur du paramètre λ

Lois usuelles

On désignera par δ_α, l'estimateur associé à la distribution réalisée d'ordre α. Nous supposerons que léchantillon est réduit à une seule observation pour faciliter les calculs (ou pour des raisons d'exhaustivité).

Exemple 2

a - Loi de Poisson : P(λ)

Elle est définie par

$$f(x/\ \lambda) = exp(-\lambda)\frac{\lambda^x}{x}\mathbf{1}_{\mathbf{N}}(x)\ ;\ \lambda > 0$$

L'α-distribution correspond à la loi de densité

$$\phi_\alpha(\lambda/\ x) = \frac{\alpha^{\alpha x+1}}{\Gamma(\alpha x+1)}\lambda^{\alpha x} exp(-\alpha\lambda)\mathbf{1}_{[0,\ +\infty[}(\lambda)$$

autrement dit

$$\phi_\alpha(\lambda/\ x) \sim G[\alpha x + 1,\ \alpha]$$

On a alors :

$$\mathbf{E}^{\phi_\alpha}(\lambda) = \frac{\alpha x + 1}{\alpha} \longrightarrow \delta_\alpha(x) = x \quad \text{lorsque} \quad \alpha \longrightarrow +\infty$$

Par conséquent, si $x_1, \ldots, x_n$ sont i.i.d de densité $f(x/\lambda)$, on obtient :

$$\delta_\alpha(x) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

b - Loi exponentielle : $Exp(\theta)$

Elle est définie par la densité :

$$f(x/\theta) = \theta e^{-\theta x} \mathbf{1}_{[0,\ +\infty[}(x) \qquad \theta > 0$$

Un calcul simple montre que l'α-distribution $\phi_\alpha(\theta/\ x)$ correspond à la densité de la loi Gamma $G[\alpha + 1,\ \alpha x]$.

On a alors :

$$\mathbf{E}^{\phi_\alpha}(\ \theta) = \frac{\alpha + 1}{\alpha x} \longrightarrow \delta_\alpha(x) = \frac{1}{x} \quad \text{lorsque} \quad \alpha \longrightarrow +\infty$$

c - Loi normale : $N[m,\ \sigma]$ (σ est supposé connu)

La densité est donnée par l'expression

$$f(x/\ m) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \qquad x \in \Re$$

On vérifie que

$$\phi_\alpha(m/\ x) \sim N[x,\ \frac{\sigma}{\sqrt{\alpha}}]$$

et que donc

$$\mathbf{E}^{\phi_\alpha}(m) = \delta_\alpha(x) = x$$

d - Loi Gamma : $G(a,\ \alpha)$

Elle est définie par la densité

$$f(x/\ a,\ \theta) = \frac{\theta^a}{\Gamma(a)} e^{-\theta x} x^{a-1} \mathbf{1}_{[0,\ +\infty[}(x) \qquad a > 0\ ;\ \theta > 0$$

L'α-distribution associée à cette loi correspond ici à une loi Gamma :

$$\phi_\alpha(\theta/\ x,\ a) \sim G[\alpha a + 1,\ \alpha x]$$

et que finalement,

$$\mathbf{E}^{\phi_\alpha}(\theta) = \frac{\alpha a + 1}{\alpha x} \longrightarrow \delta_\alpha(x) = \frac{a}{x} \quad \text{lorsque} \quad \alpha \longrightarrow +\infty$$

Nous avons donc tenter de proposer ici, une nouvelle approche de la distribution généralisée d'ordre α. En effet, les mesures informationnelles probabilistes d'ordre α (ou de type α), nous ont permis d'introduire des familles de lois d'ordre α, dites α-distributions. On montre que lorsque l'on se restreint à la famille exponentielle, l'α-distribution peut être utilisée pour la détermination de l'estimateur du maximum de vraisemblance, quand l'ordre α devient infiniment grand. Ainsi, lorsque le calcul analytique de l'estimateur s'avère délicat, on peut contourner cette difficulté par la mise au point de la procédure fondée sur ce résultat.

$\triangle$

Chapitre 7

Conclusion et perspectives

7.1 Conclusion

Nous avons, dans cette thèse, tenté d'utiliser une statistique construite à partir d'une mesure d'information et d'une distribution généralisée ϕ_α, lorsque le modèle de base retenu est un modèle paramétrique.

Initialement, l'α-distribution avait été introduite par Hammad, lorsque $\alpha \in [1, \infty[$, α non nécessairement entier, en vue d'envisager un parallèlisme entre α et le temps $t \in \mathbf{R}_+$ relatif à un processus stochastique $X(t)$ à temps continu, avec la correspondance $t = \dfrac{1}{\alpha - 1}$, $\alpha \in [1, \infty[$ ou bien $\alpha = 1 + \dfrac{1}{t}$

a - Dans un premier temps, nous avons privilégié une autre approche de cette distribution d'ordre α, en vue d'une inférence statistique (sur le paramètre), en restreignant la valeur de α au cas $\alpha = 1$, correspondant à la distribution de probabilité ordinaire.

En conséquence, nous avons montré que la statistique de la divergence $\widehat{J}_n$ est asymptotiquement équivalente, sous l'hypothèse nulle, aux statistiques de test classiques. On constate que la seule différence engendrée par ce rapprochement n'est observable qu'en distance finie.

La simulation par Monte Carlo, d'un modèle de régression d'une part et d'une

distribution exponentielle d'autre part, a fait l'objet d'un chapitre particulier, car l'évaluation constitue une étape importante notamment dans l'optique d'une décision statistique.

Ainsi, il ressort de cette comparaison, que la statistique de test $\widehat{J}_n$ que nous avons proposé apparaît comme étant plus robuste que ces concurrents habituels, dans la mesure où elle reste plus stable par rapport aux fluctuations de l'hypothèse alternative, comme on a pu le voir dans le cas où la loi des observations admet une densité exponentielle.

Nous retiendrons qu'il n'existe pas en l'occurrence de test optimal, en ce sens qu'on ne peut trouver un test qui soit plus puissant que tous les autres.

Enfin, à part la facilité de calcul découlant des estimateurs utilisés (maximun de vraisemblance, maximisation non contraint, optimisation d'une fonction objectif ...), il n'existe pas de critère permettant d'obtenir le meilleur choix, à priori entre ces différentes statistiques de test classiques. On peut donc considérer, que la procédure de test, fondé sur $\widehat{J}_n$ que nous avons proposée, peut être interprétée comme un compromis entre Wald et le Rapport de vraisemblance, dans la mesure où, elle est assimilable à une valeur moyenne de ces statistiques, lorsque l'on suppose que le modèle retenu suit une loi exponentielle.

b - Dans un deuxième temps, nous avons considéré $\alpha \in \mathbf{R}_+$, et nous avons proposé une approche numérique à partir d'un algorithme basé sur des techniques de calculs bayésiens, pour la résolution des équations de vraisemblance, lorsqu'aucun calcul analytique n'est possible.

A notre connaissance, aucun travail n'a encore abordé les situations suivantes :

- une recherche sur l'aspect comparatif entre une statistique d'une mesure d'information et les statistiques de test classique habituellement utilisés en dimension finie,
- l'étude et l'interprétation de ϕ_α sous l'approche bayésienne.

Ces deux points constituent ainsi notre principale contribution.

7.2 Perspectives

Nous avons pu constater un certain nombre de limites dans ce travail. Elles sont surtout relatives à l'interprétation de l'ordre α. Une tentative de rapprochement avec la théorie de l'échantillonnage n'a pas donné tous les résultats escomptés. En effet, l'α-distribution n'a pas pu être exploitée davantage, notamment lorsque ce paramètre est entier. Cependant il serait souhaitable, d'approfondir le rôle que pourrait jouer α (lorsque $\alpha \in \mathbf{N}^* - 1$), pour au moins une famille de distribution, comme par exemple, les lois à paramètre de position dont la densité s'écrit sous la forme "$f(x-\theta)$". C'est le cas par exemple pour :

(1) la loi normale (la variance étant supposée connue),

(2) la loi exponentielle (le paramètre d'échelle est connu),

(3) la loi de Pearson de Type III (le paramètre de position étant seul inconnu).

Dans les cas particuliers (1) et (2), $\phi_\alpha(x/\theta)$ coincide avec la densité d'une statistique exhaustive, ce qui justifie un éventuel usage pour une inférence sur le paramètre d'intérêt. En effet, on a :

- si $X \sim N[\theta, \sigma^2]$ alors $\phi_\alpha(x/\theta)$ est la loi de la moyenne $\bar{X}_\alpha$

- si $X \sim Exp(\lambda, \theta)$, dans ce cas $\phi_\alpha(x/\theta)$ (le paramètre d'échelle est supposé connu), représente la loi de la statistique
exhaustive $T = Inf X_i$

Une deuxième perspective soulignée dans cette thèse résulte de l'utilisation d'autres mesures d'information (les (h,ϕ)-divergences, l'entropie, etc...), dans le cadre de la théorie de l'estimation ou des tests, en vue d'une étude comparative avec les méthodes usuelles, lorsqu'on se restreint aux petits échantillons.

$\triangle$

Chapitre 8

Annexes

8.1 Annexe 1

Preuve du théorème 3 :

Posons:

$$\psi(\theta) = \Delta_r[f,\, h(\theta)]$$

Le développement de Taylor autour de θ donne:

$$\psi(\widehat{\theta}) = \psi(\theta) + \sum_{i=1}^{k}(\widehat{\theta}_i - \theta_i)\frac{\partial\psi\theta)}{\partial\theta_i} + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}(\widehat{\theta}_i - \theta_i)\frac{\partial^2\psi(\theta)}{\partial\theta_i\partial\theta_j}(\widehat{\theta}_j - \theta_j) + R_n$$

si $f_i = h_i(\theta)$; $i = 1,2,\ldots,M$, on obtient alors : $\dfrac{\partial\psi\theta)}{\partial\theta} = \psi(\theta) = 0$

Soit :

$$\psi(\widehat{\theta}) = \psi(\theta)\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}(\widehat{\theta}_i - \theta_i)\frac{\partial^2\psi(\theta)}{\partial\theta_i\partial\theta_j}(\widehat{\theta}_j - \theta_j) + R_n$$

$$\frac{n}{r}\psi(\widehat{\theta}) = \sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{l=1}^{M} n(\widehat{\theta}_i - \theta_i)\frac{1}{h_l(\theta)}\frac{\partial h_l(\theta)}{\partial\theta_i}\frac{\partial h_l(\theta)}{\partial\theta_j}(\widehat{\theta}_j - \theta_j) + R'_n$$

En posant :

$$\begin{aligned} I(\theta) &= \sum_{l=1}^{M}\frac{1}{h_l(\theta)}\frac{\partial}{\partial\theta_i}h_l(\theta)\frac{\partial}{\partial\theta_j}h_l(\theta) \\ &= \sum_{l=1}^{M}\frac{\partial}{\partial\theta_i}h_l(\theta)\frac{1}{h_l(\theta)}\frac{\partial}{\partial\theta_j}h_l(\theta) \end{aligned}$$

on obtient :

$$\frac{n}{r}\Delta_r[f,\, h(\widehat{\theta})] \longrightarrow \chi^2_k$$

8.2 Annexe 2

Preuve du théorème 4 :

Considérons toujours la fonction $\psi(\theta) = \Delta_r[f,\, h(\theta)]$

Le développement de Taylor autour de θ à l'ordre 1, donne :

$$\psi(\widehat{\theta}) = \psi(\theta) + \sum_{i=1}^{k}(\widehat{\theta_i} - \theta_i)\frac{\partial\psi\theta)}{\partial\theta_i} + R_n$$

Or :

$$\begin{aligned}\frac{\partial\psi\theta)}{\partial\theta} &= \frac{1}{r-1}\sum_{l=1}^{M}\Big((1-r)\frac{f_l^r}{h_l^r(\theta)} + r\frac{h_l^{r-1}(\theta)}{f_l^{r-1}}\Big)\frac{\partial}{\partial\theta_i}h_l(\theta)\\ &= m(r)p_i(\theta)\end{aligned}$$

avec :

$$m(r) = \frac{1}{r-1}\sum_{l=1}^{M}\Big((1-r)\frac{f_l^r}{h_l^r(\theta)} + r\frac{h_l^{r-1}(\theta)}{f_l^{r-1}}\Big) \qquad et \qquad p_i(\theta) = \frac{\partial}{\partial\theta_i}h_l(\theta)$$

Ce qui donne :

$$\sqrt{n}\{\psi(\widehat{\theta}) - \psi(\theta)\} = m(r)\sqrt{n}(\widehat{\theta} - \theta)p(\theta) + R_n$$

On peut donc dire que les variables aléatoires $\sqrt{n}\{\psi(\widehat{\theta}) - \psi(\theta)\}$ et $m(r)\sqrt{n}(\widehat{\theta} - \theta)p(\theta)$ ont la même distribution asymptotique, à savoir :

$$\sqrt{n}\{\psi(\widehat{\theta}) - \psi(\theta)\} \longrightarrow N[0,\, \Gamma^2]$$

avec :

$$\Gamma^2 = \lambda^2\sum_{i,j}I_{ij}(\theta)\Big\{\sum_{l=1}^{M}\Big((1-r)\frac{f_l^r}{h_l^r(\theta)} + r\frac{h_l^{r-1}(\theta)}{f_l^{r-1}}\Big)\Big\}^2\frac{\partial}{\partial\theta_i}h_l(\theta)\frac{\partial}{\partial\theta_j}h_l(\theta)$$

où $\lambda = (r-1)^{-1}$ et où I_{ij} représente le terme général de la matrice d'information de Fisher.

8.3 Annexe 3

Preuve du théorème 5 :

L'expression de la variance Σ^2 associée à la statistique $\widehat{D}_n$, est déterminée moyennant un développement limité de Taylor à l'ordre 1 des fonctions $\hat{\Delta}_{1/2}[f, h(\theta)]$ et $\hat{\Delta}_{1/2}[f, g(\pi)]$.

Posons :

$$\hat{\Delta}_{1/2}(\theta) = \hat{\Delta}_{1/2}[f, h(\theta)] \quad ; \quad \hat{\Delta}_{1/2}(\pi) = \hat{\Delta}_{1/2}[f, g(\pi)]$$

$$\frac{\partial}{\partial\theta}\Delta_{1/2}(\theta) = p(\theta) \quad et \quad \frac{\partial}{\partial\pi}\Delta_{1/2}(\pi) = q(\pi)$$

On obtient alors :

$$\sqrt{n}\hat{\Delta}_{1/2}(\theta) = \sqrt{n}\Delta_{1/2}(\theta) + \sqrt{n}\sum_i(\hat{\theta}_i - \theta_i)\frac{\partial}{\partial\theta_i}\Delta_{1/2}(\theta) + R_n^1 \qquad (8.1)$$

$$\sqrt{n}\hat{\Delta}_{1/2}(\pi) = \sqrt{n}\Delta_{1/2}(\pi) + \sqrt{n}\sum_i(\hat{\pi}_i - \pi_i)\frac{\partial}{\partial\pi_i}\Delta_{1/2}(\pi) + R_n^2 \qquad (8.2)$$

Par différence des relations (8.1) et (8.2) :

$$\begin{aligned}\widehat{D}_n &= D_n + \sqrt{n}\sum_i(\hat{\theta}_i - \theta_i)\, p_i(\theta) - \sqrt{n}\sum_i(\hat{\pi}_i - \pi_i)\, q_i(\pi) + R_n \\ &= D_n + (p^t(\theta)\ ,\ -q^t(\pi))\begin{pmatrix}\sqrt{n}(\hat{\theta} - \theta) \\ \sqrt{n}(\hat{\pi} - \pi)\end{pmatrix} + R_n \qquad (8.3)\end{aligned}$$

En posant :

$$C^t = C^t(\theta, \pi) = (p^t(\theta)\ ,\ -q^t(\pi)) \quad ; \quad \hat{\eta} = \sqrt{n}(\hat{\theta} - \theta)$$

$$et \quad \hat{\delta} = \sqrt{n}(\hat{\pi} - \pi)$$

on obtient :

$$\widehat{D}_n = D_n + C^t(\theta\ ,\ \pi)\begin{pmatrix}\hat{\eta} \\ \hat{\delta}\end{pmatrix} + R_n \qquad (8.4)$$

et comme $R_n = R_n^1 - R_n^2 \to 0$ quand n tend vers l'infini, on en déduit que les deux variables aléatoires $\widehat{D}_n - D_n$ et $C^t(\theta\ ,\ \pi)\begin{pmatrix}\hat{\eta} \\ \hat{\delta}\end{pmatrix}$ ont asymptotiquement la même

distribution, autrement dit :

$$\widehat{D}_n - D_n \to N[0, \Sigma^2]$$

avec

$$\Sigma^2 = C^t \Lambda C$$

où

$$\Lambda = \begin{bmatrix} I^{-1}(\theta) & \mathbf{E}(\widehat{\eta}\ \widehat{\delta}^t) \\ \mathbf{E}(\widehat{\delta}\ \widehat{\eta}^t) & I^{-1}(\pi) \end{bmatrix}$$

puisque $\widehat{\eta}$ et $\widehat{\delta}$ sont des variables aléatoires centrées.

$\triangle$

Bibliographie

[1] H. Akaike : "Information theory and Extension of the Likelihood Ratio Principe", *Proceedings of the second International Symposium of Information theory , ed. By. Pietrov, B.N and Csaki, F. Budapest: Akademiai Kiado, pp 257-281 (1973)*

[2] S. I. Amari : "Differential-Geometric Methods in Statistic ", *Lecture Notes in Statistics, Spring Verlag, Berlin (1985)*

[3] S. I. Amari : "Differential Geometry of Statistics : Towards new Developments", *In : NATO Workshop on Differential Geometry in Statistical Inference, London, 9-11 April. (1984)*

[4] A. Bhattacharyya : "On a measure of divergence between two statistical populations defined by their probability distributions", *Bull. Calcutta Math.Soc., 35, 99-109, (1943)*

[5] T. S. Breush and A. R. Pagan : "The Lagrange Multiplier Test and its applications to model Specification in Econometrics ", *Review of Economic Studies 47 pp 239-253 (1980)*

[6] J. Burbea and C. R. Rao : "On the convexity of some Divergence Measures Based on Entropy Functions ", *IEEE Trans. on Information Theory, IT - 28, 489-495 (1982b)*

[7] D. R. Cox and D. V. Hinkley : "Theoretical Statistics ", *London, Chapman and Hall (1974)*

[8] I. Csiszar : "Information-type measures of difference of probability distributions and indirect observations ", *Studia Sci. Math. Hung. 299-318 (1967)*

[9] R. Davidson and J.G. Mackinnon: "Estimation and Inference in Econometrics ", *New York Oxford, Oxford University Press (1993)*

[10] R. Davidson and J.G. Mackinnon: "Graphical methods for investigating the size and Power of hypothesis tests ", *Documents de travail G.R.E.Q.A.M n94A23 Juin (1994).*

[11] H. T. Davis: "The theory of econometrics ", *The Principia Press, Bloomington, IN (1941)*

[12] L. Devroye: "Non-uniform random variate generation ", *Springer-Verlag, New York (1985)*

[13] C. Gourieroux et A. Monfort: "Statistique et modèles économétriques ", *2 édition Economica Paris (1996)*

[14] P. Hammad: "Information, Systèmes et distributions ", *Editions Cujas, Paris (1987)*

[15] P. Hammad: "Information Theory, Statistical Decision Functions, Random Processus ", *Transactions of the Eleventh Prague Conference - Prague, from August 27 to 31, 1990.*
Academia Publishing House of the Czechoslovak Academy of Sciences Prague (1992)

[16] P. Hammad et P. Ngom: "Test d'ajustement et test de choix fondés sur une distance informationnelle généralisée ", *Rev. Stat. Appl. (1988) (à paraître)*

[17] R. V. L. Hartley: "Transmission of Information ", *Bell System Tech. Journal, 7, 535 (1928)*

[18] D. F. Hendry and J. F. Richard: "The Econometric Analysis of Economic Time series ", *International Statistical Review 51 (2) pp. 111-163 (1983)*

[19] E. T. Jaynes: "Information Theory and statistical mechanics ", *Phys. Rev. Vol 106 pp. 620-630 (1957)*

[20] H. Jeffreys: "An Invariant form of the Prior Probability in Estimation Problems ", *Proc. Royal Soc. Ser. A, 186, 453-561 (1946)*

[21] H. Jeffreys: "Theory of probability ", *(3rd. rev. ed.), Oxford University Press, London (1967)*

[22] M. G. Kendall and A. Stuart: "The advanced theory of statistics ", *Ch. Griffin and Co. London, (1969)*

[23] S. Kullback and M. Leibler: "On the information and sufficiency ", *Ann. Math. Statist. 27, 986-1005 (1951)*

[24] S. Kullback: "Information theory and statistics ", *Dover Plublications, Inc. New York (1968)*

[25] E. Lehmann: "Theory of Point Estimation ", *Wiley New York (1983)*

[26] E. Lehmann: "Testing Statistical Hypothesis ", *Wiley New York (1986)*

[27] E. Maasoumi: "Information theory ", *Vol. 2, New York (1988b) Stockton Press, 846-51 . Reprinted in New Palgrave: Econometrics, Norton (1990)*

[28] K. Matusita: "On theory of décision functions", *Ann. Inst. Statist. Math., 3, 17-35 (1951)*

[29] K. Matusita: "On the notion of affinity of several distributions and some of its applications", *Ann. Inst. Statist. Math., 19, 181-192 (1967)*

[30] D.Morales, L.Pardo, M.Salicrù and M.L Menendez: "A test of independance based on the (r, s)-directed divergence", *Tamkang Journal of Mathematics, Vol.23, N2, Summer (1992)*

[31] D.Morales, L.Pardo, M.Salicrù and M.L Menendez: "Some statistical applications of (r, s)-directed divergences ", *Utilit. Mathemat. 42, 115-127 (1992)*

[32] D.Morales, L.Pardo, M.Salicrù and M.L Menendez: "Asyptotics properties of (r, s)-directed divergence in a stratified sampling ", *Applied Mathematics and computation, 15, 131-152 (1993)*

[33] D.Morales, L.Pardo, M.Salicrù and M.L Menendez: "The ϕ–divergence statistics in bivariate multinomial populations including stratification ", *Metrika, 40, 223-235 (1993)*

[34] D.Morales, L.Pardo, M.Salicrù and M.L Menendez: "Asymptotic properties of divergence statistics in a stratified random sampling and its applications to test statistical hypotheses", *Journal of Statitical Planning and Inference, 38, p.201-222 North-Holland (1994).*

[35] D.S Moore: " Chi-Squared Tests", *in statudies in statistics, ed. by HoGG , R.V. Volume, The Mathemical Association of America. (1978)*

[36] D.S Moore: "Test of Chi-Squared type", *Goodness-of-fit techniques, ed. D'Agostino, R.B and Stephens, M.A (1986)*

[37] A.M Mathai and P.N Rathie: "Basic Concepts of Information Theory and Statistics", *Wiley, New York (1975)*

[38] K. Pearson: "On the criterion that a given System of deviation from the probable in the case of a correlated System of Variables is Such that it can be reasonably supposed to have Arisen from Random Sampling", *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science , 50, pp 157-175 (1990)*

[39] A. Renyi: "Calcul des probabilités (avec un appendice sur la théorie de l'information)", *Dunod, Paris (1966)*

[40] A. Renyi: "On measures of entropy and Information ", *Proced. 4th Berkeley Symp. Math. Statist. and Prob. 1, 547-561 U. of California Press (1961)*

[41] C. Robert: "L'analyse statistique bayésienne ", *Economica Paris (1992)*

[42] R. J. Serfling: "Approximation theorem of Mathematical Statistics ", *Wiley (1980)*

[43] B. D. Sharma and D. P. Mittal: "New nonadditive measures of entropy for discrete probability distributions", *J. Math. Sci., 10, 28-40 (1977)*

[44] Shannon. C: "A mathematical theory of comunications ", *Bell System Tech. J. 27, 379-423 (1948)*

[45] I.J Taneja: "Some Contributions to Information Theory I (A survey): On Measures of Information", *J. Comb., Inform. Sys. Sci. 4(4), 253-274. (1979)*

[46] I.J Taneja: "On generalized information measures and their applications ", *Adv. Elect. and Elect. Phis. 76, 327-413 (1989)*

[47] : "Theory of statistical Inference and Information ", ***Kluwer** Academic Press Publishers, Dordrecht (1989)*

[48] Q.H Vuong: "Likelihood Ratio tests for model Selection and non-nested Hypotheses", *Econometrica, 57, pp 257-306 (1989)*

[49] Q.H Vuong and W. Wuang: "Selecting Estimated Models using Chi-Square Statistics", *Annales d'économie et de Statistique, 30, pp 143-164 (1993)*

[50] G. S. Watson : "Some Recent Results in Chi-Square Goodness-of-Fit Tests", *Biometrics, 15, pp. 440-468 (1959) 30, pp 143-164 (1993)*

[51] A. M. Yaglom et I. M. Yaglom : "Probabilité et Information ", *Dunod Paris (1959)*